AD-A254 755

N DOCUMENT

PHOTOGRAPH THIS SHEET

DTIC ACCESSION NUMBER

LEVEL

(1)

INVENTORY

AFOSR-TR-92-0709

DOCUMENT IDENTIFICATION

25 Jun 92

Approved for public release;
Distribution Unlimited

DISTRIBUTION STATEMENT

ACCESSION FOR		
NTIS	GRA&I	☑
DTIC	TRAC	☐
UNANNOUNCED		☐
JUSTIFICATION		
BY		
DISTRIBUTION/		
AVAILABILITY CODES		
DISTRIBUTION	AVAILABILITY AND/OR SPECIAL	
A-1		

DISTRIBUTION STAMP

DTIC
ELECTE
AUG 3 1992
S C

DATE ACCESSIONED

HANDLE WITH CARE

DTIC QUALITY INSPECTED 5

DATE RETURNED

92 7 2

DATE RECEIVED IN DTIC

92-20721

REGISTERED OR CERTIFIED NUMBER

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM JUN 90 70A

DOCUMENT PROCESSING SHEET

PREVIOUS EDITIONS MAY BE USED UNTIL STOCK IS EXHAUSTED.

LOAN DOCUMENT

AD-A254 755

UNITED STATES AIR FORCE

1990 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-88-C-0053

RESEARCH REPORTS

VOLUME III OF IV

DTIC
ELECTE
AUG 3 1992
S C D

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

June 1992

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

92-20721

92 7 30 052

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
25 Jun 92

3. REPORT TYPE AND DATES COVERED
ANNUAL 1 Jan 91 to 31 Dec 91

4. TITLE AND SUBTITLE
USAF 1990 Research Initiation Program Conducted by Universal Energy Systems, Inc VOL #

5. FUNDING NUMBERS
F49620-88-C-0053
2305/D5

6. AUTHOR(S)
Rod Darrah

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Universal Energy Systems, Inc.
Dayton OH

8. PERFORMING ORGANIZATION REPORT NUMBER
9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
Bolling AFB DC

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
(U)

12b. DISTRIBUTION CODE
(U)

13. ABSTRACT (Maximum 200 words)

This program is for follow-on research efforts for the participants in the Summer Faculty Research Program. Funding is provided to establish RIP awards to about half the number of participants in the SFRP. Participants in the 1990 SFRP competed for funding under the 1990 RIP. Evaluation of the proposals were made by the contractor. Evaluation criteria consisted of: 1. Technical excellence of the proposal 2. Continuation of the SFRP effort 3. Cost sharing by the university. The list of proposals selected for award was forwarded to AFOSR for approval of funding and for research efforts to be completed by 31 December 1991. The following summarizes the events for the evaluation of proposals and award of funding under the RIP. A. RIP proposals were submitted to the contractor by 1 November 1991. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university. B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories. C. Subcontracts were negotiated with the Universities. There were a total of 92 RIP awards made under the 1990 program.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
(U)

18. SECURITY CLASSIFICATION OF THIS PAGE
(U)

19. SECURITY CLASSIFICATION OF ABSTRACT
(U)

20. LIMITATION OF ABSTRACT
(U)

UNITED STATES AIR FORCE

1990 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-88-C-0053

RESEARCH REPORTS

VOLUME III OF IV

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

June 1992

TABLE OF CONTENTS

SECTION	PAGE
INTRODUCTION	ii
STATISTICS	iii
PARTICIPANT LABORATORY ASSIGNMENT	vii
RESEARCH REPORTS	xv

INTRODUCTION

Research Initiation Program - 1990

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially, this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1990 SFRP competed for funding under the 1990 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1990, following their participation in the 1990 SFRP.

Evaluation of these proposals were made by the contractor. Evaluation criteria consisted of:

1. Technical excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the university

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1991.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. RIP proposals were submitted to the contractor by 1 November 1990. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1990 and December 1991.

Copies of the final reports are presented in Volumes I through IV of the 1990 Research Initiation Program Report. There were a total of 92 RIP awards made under the 1990 program.

STATISTICS

PROGRAM STATISTICS

Total SFRP Participants	165
Total RIP Proposals submitted by SFRP	129
Total RIP Proposals submitted by GSRP	7
Total RIP Proposals submitted	136
Total RIP's funded to SFRP	88
Total RIP's funded to GSRP	4
Total RIP's funded	92
Total RIP Proposals submitted by HBCU's	9
Total RIP Proposals funded to HBCU's	2

LABORATORY PARTICIPATION

Laboratory	Participants	Submitted	Funded
AAMRL	10	6	5
WRDC/APL	11	9	7
ATL	11	9 (1 GSRP)	8 (1 GSRP)
AEDC	7	10 (3 GSRP)	6 (GSRP)
WRDC/AL	9	8 (1 GSRP)	6
ESMC	0	0	0
ESD	1	1	0
ESC	11	8	5
ETL	3	2	1
WRDC/FDL	11	11	6
FJSRL	8	6	4
AFGL	11	9 (1 GSRP)	6
HRL	14	12	8
WRDC/ML	11	8	6
OEHL	4	2	1
AL	7	5	4
RADC	12	9	6
SAM	17	14	8
WL	6	7 (1 GSRP)	5 (1 GSRP)
WHMC	1	0	0
Total	165	136	92

LIST OF PARTICIPATING UNIVERSITIES

Alabama, University of	-	4
Alfred University	-	1
Arizona State University	-	1
Brigham Young University	-	1
Butler University	-	1
California State University	-	1
Cincinnati, University of	-	1
Colorado-Colorado Springs, Univ. of	-	2
Denver, University of	-	1
Duke University	-	1
Fairleigh Dickinson University	-	1
Florida Institute of Tech.	-	1
Florida University of	-	1
Houston, University of	-	1
Houston-Victoria, Univ. of	-	1
Idaho State University	-	1
Indiana University	-	1
Indiana-Purdue University	-	1
Kansas State University	-	1
Kent State University	-	1
Kentucky, University of	-	2
Lowell, University of	-	1
MIT	-	1
Maine, University of	-	2
Marshall University	-	1
Maryland-Baltimore, University of	-	1
Memphis State University	-	2
Michigan Tech. University	-	1
Minnesota, University of	-	1
Mississippi State University	-	3
Missouri-Kansas City, Univ. of	-	1
Missouri-Rolla, University of	-	1
Morehouse College	-	1
New Orleans, University of	-	1
New York, City College of	-	1
New York-Buffalo, State Univ. of	-	1
North Dakota State University	-	2
North Texas, University of	-	2
Northwestern University	-	2
Ohio State University	-	2
Oklahoma State University	-	1
Pennsylvania State University	-	1
Puerto Rico, University of	-	1
Purdue Calumet University	-	1
Rensselaer Polytechnic Univ.	-	1
Rhode Island, University of	-	1
Scranton, University of	-	2
South Dakota, University of	-	1
South Florida, University of	-	1
Southern Illinois University	-	1
Southwest Texas State Univ.	-	1
Syracuse University	-	1
Texas A&M University	-	1
Texas Tech. University	-	1
Texas-Arlington, Univ. of	-	1
Texas-Austin, Univ. of	-	1
Texas-San Antonio, Univ. of	-	2
Trinity University	-	1
Tufts University	-	1
Utah State University	-	2
Utica College	-	1
Vanderbilty University	-	3
Virginia Poly. Instit. State Univ.	-	1
Washington State University	-	2
Wellesley College	-	1
West Texas State University	-	1
West Virginia University	-	1
Worchester Polytechnic Institute	-	1
Wright State University	-	4
Wyoming, University of	-	1
Total	-	92

PARTICIPANTS LABORATORY ASSIGNMENT

AERO PROPULSION AND POWER DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Muhammad Choudhry
West Virginia University
Specialty: Electrical Engineering

Dr. Donald Dareing
University of Florida
Specialty: Mechanical Engineering

Dr. Paul Hedman
Brigham Young University
Specialty: Chemical Engineering

Dr. K. Sankara Rao
North Dakota State University
Specialty: Electrical Engineering

Dr. Larry Roe
Virginia Poly. Instit. State University
Specialty: Mechanical Engineering

Dr. Kaveh Tagavi
University of Kentucky
Specialty: Mechanical Engineering

Dr. Richard Tankin
Northwestern University
Specialty: Mechanical Engineering

ARMAMENT DIRECTORATE
(Eglin Air Force Base)

Dr. Charles Camp
Memphis State University
Specialty: Civil Engineering

Dr. Arnold Carden
University of Alabama
Specialty: Metallurgy

Dr. Charles Fosha
University of Colorado
Specialty: Electrical Engineering

Dr. John George
University of Wyoming
Specialty: Applied Mathematics

Dr. Kevin Moore
Idaho State University
Specialty: Electrical Engineering

Dr. William Siuru
University of Colorado
Specialty: Mechanical Engineering

Dr. Kenneth Sobel
City College of New York
Specialty: Electrical Engineering

Mr. Randy Gove (GSRP)
University of Alabama
Specialty: Physics

ARMSTRONG LABORATORY
(Brooks Air Force Base)

Dr. Phillip Bishop
University of Alabama
Specialty: Exercise Physiology

Dr. Robert Blystone
Trinity University
Specialty: Zoology

Dr. Bruno Breitmeyer
University of Houston
Specialty: Experimental Psychology

Dr. Vito DelVecchio
University of Scranton
Specialty: Biochemistry

Dr. Pushpa Gupta
University of Maine
Specialty: Mathematics

Dr. Ramesh Gupta
University of Maine
Specialty: Mathematical Statistics

Dr. John Szarek
Marshall University
Specialty: Pharmaceutical

Dr. Steven Waller
University of South Dakota
Specialty: Pharmacology

ARNOLD ENGINEERING DEVELOPMENT CENTER
(Arnold Air Force Base)

Dr. Carlyle Moore
Morehouse College
Specialty: Physics

Dr. Olin Norton
Mississippi State University
Specialty: Mechanical Engineering

Dr. Richard Peters
Vanderbilt University
Specialty: Electrical Engineering

Dr. Chun Su
Mississippi State University
Specialty: Physics

Mr. Ben Abbott (GSRP)
Vanderbilt University
Specialty: Computer Engineering

Mr. Theodore Bapty (GSRP)
Vanderbilt University
Specialty: Computer Engineering

AVIONICS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Michael Breen
Alfred University
Specialty: Mathematics

Dr. R. H. Cofer
Florida Institute of Tech.
Specialty: Electrical Engineering

Dr. Lawrence Hall
University of South Florida
Specialty: Computer Science

Dr. Kevin Kirby
Wright State University
Specialty: Computer Science

Dr. Richard Miers
Indiana Univ. - Purdue Univ.
Specialty: Physics

Mr. Eric Byrne (GSRP)
Kansas State University
Specialty: Computer Science

CREW SYSTEMS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Richard Backs
Wright State University
Specialty: Psychology

Mr. John Duncan
Kent State University
Specialty: Technology

Dr. Martin Hagan
Oklahoma State University
Specialty: Electrical Engineering

Dr. Ashok Krishnamurthy
Ohio State University
Specialty: Electrical Engineering

Dr. Amit Patra
University of Puerto Rico
Specialty: Mechanical Engineering

ELECTRONIC TECHNOLOGY LABORATORY

Dr. Ashok K. Goel
Michigan Tech. University
Specialty: Electrical Engineering

ENGINEERING AND SERVICES CENTER
(Tyndall Air Force Base)

Dr. William Bannister
University of Lowell
Specialty: Organic Chemistry

Dr. Joseph Dreisbach
University of Scranton
Specialty: Chemistry

Dr. Michael McFarland
Utah State University
Specialty: Biological Engineering

Dr. Perry McNeill
University of North Texas
Specialty: Education

Dr. George Veyera
University of Rhode Island
Specialty: Civil Engineering

FLIGHT DYNAMICS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Marvin Hamstad
University of Denver
Specialty: Solid Mechanics

Dr. Chin Hsu
Washington State University
Specialty: Electrical Engineering

Dr. David Hui
Worcester Poly. Institute
Specialty: Aerospace Engineering

Dr. Yulian Kin
Purdue Calumet
Specialty: Mechanical Engineering

Dr. Byung-Lip Lee
Pennsylvania State University
Specialty: Materials Science

Dr. Lawrence Zavodney
Ohio State University
Specialty: Mechanical Engineering

FRANK J. SEILER RESEARCH LABORATORY
(United States Air Force Academy)

Dr. Theodore Burkey
Memphis State University
Specialty: Chemistry

Dr. Richard Carlin
University of Alabama
Specialty: Chemistry

Dr. Ephrahim Garcia
State Univ. of New York-Buffalo
Specialty: Aerospace Engineering

Dr. Thomas Posbergh
University of Minnesota
Specialty: Electrical Engineering

GEOPHYSICS DIRECTORATE
(Hanscom Air Force Base)

Dr. Chia-Bo Chang
Texas Tech. University
Specialty: Meteorology

Dr. George Jumper
Worcester Poly. Institute
Specialty: Mechanical Engineering

Dr. Charles Lishawa
Utica College
Specialty: Physical Chemistry

Dr. Craig Rasmussen
Utah State University
Specialty: Physics

Dr. Glenn Stark
Wellesley College
Specialty: Physics

Dr. John Wills
Indiana University
Specialty: Physics

HUMAN RESOURCES DIRECTORATE
(Brooks, Williams and Wright-Patterson Air Force Base)

Dr. Margaret Batschelet
University of Texas-San Antonio
Specialty: English

Dr. James Dykes
University of Texas-San Antonio
Specialty: Psychology

Dr. Verlin Hinsz
North Dakota State University
Specialty: Psychology

Dr. Delayne Hudspeth
University of Texas-Austin
Specialty: Education

Dr. Gillray Kandel
Rensselaer Poly. Institute
Specialty: Experimental Psychology

Dr. William Moor
Arizona State University
Specialty: Industrial Engineering

Dr. Joan Rentsch
Wright State University
Specialty: Industrial Psychology

Dr. Stanley Stephenson
Southwest Texas State University
Specialty: Psychology

MATERIALS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. John Connolly
University of Missouri-Kansas
Specialty: Chemistry

Dr. David Grossie
Wright State University
Specialty: Chemistry

Dr. Prasad Kadaba
University of Kentucky
Specialty: Physics

Dr. Joseph Lambert
Northwestern University
Specialty: Chemistry

Dr. Martin Schwartz
University of North Texas
Specialty: Physical Chemistry

Dr. Hai-Lung Tsai
University of Missouri-Rolla
Specialty: Mechanical Engineering

OCCUPATIONAL AND ENVIRONMENTAL HEALTH DIRECTORATE
(Brooks Air Force Base)

Dr. Miguel Medina
Duke University
Specialty: Water Resources

ROCKET PROPULSION DIRECTORATE
(Edwards Air Force Base)

Dr. Shannon Lieb
Butler University
Specialty: Physical Chemistry

Dr. Thomas Pollock
Texas A&M University
Specialty: Materials Science

Dr. Hung Vu
California State University
Specialty: Applied Mechanics

Dr. Trevor Williams
University of Cincinnati
Specialty: Control Theory

ROME LABORATORIES
(Griffiss Air Force Base)

Dr. Gary Craig
Syracuse University
Specialty: Electrical Engineering

Dr. Frances Harackiewicz
Southern Illinois University
Specialty: Electrical Engineering

Dr. Shietung Peng
University of Maryland-Baltimore
Specialty: Computer Science

Dr. Daniel Ryder
Tufts University
Specialty: Chemical Engineering

Dr. Behrooz Shirazi
Southern Methodist University
Specialty: Computer Science

Dr. Wayne Smith
Mississippi State University
Specialty: Computer Science

WEAPONS DIRECTORATE
(Kirtland Air Force Base)

Dr. Gene Carlisle
West Texas State University
Specialty: Inorganic Chemistry

Dr. William Cofer
Washington State University
Specialty: Civil Engineering

Dr. Johanna Schruben
University of Houston-Victoria
Specialty: Mathematics

Dr. Peter Walsh
Fairleigh Dickinson University
Specialty: Physics

Mr. Michael Houts (GSRP)
Massachusetts Inst. of Tech.
Specialty: Nuclear Engineering

RESEARCH REPORTS

MINI-GRANT RESEARCH REPORTS

Technical Report Number	Title and Mini-Grant Number	Professor
Volume I		
Rome Laboratories		
1	Resource Management in a Heterogeneous Distributed Systems 210-11MG-107	Dr. Gary Craig
2	Analysis of Magnetically Controllable Microstrip Antennas on Ferrite Substrates 210-11MG-120	Dr. Frances Harackiewicz
3	Architectural Design and Simulation of a Parallel Signal Processor for Adaptive Space-Time Processing 210-11MG-103	Dr. Shietung Peng
4	Thermal Processing of Pb-Modified Bi-Sr-Ca-Cu-O Superconducting Thin Films Prepared by the Metalorganic Deposition (MOD) Method 210-11MG-011	Dr. Daniel Ryder
5	Parallel Architectures for AI and Dynamic Knowledge-Base Systems 210-11MG-064	Dr. Behrooz Shirazi
6	Markov Models for Simulating Error Patterns on Data Communications Links 210-11MG-012	Dr. Wayne Smith
Arnold Engineering Development Center		
7	Performance Measurement and Optimization of Parallel Systems 210-11MG-106	Mr. Ben Abbott (GSRP)
8	An Environment for Reconfigurable Parallel Processing Software 210-11MG-104	Mr. Theodore Bapty (GSRP)
9	Feasibility of Measuring Pulsed X-Ray Spectra using Photoactivation of Nuclear Isomers 210-11MG-056	Dr. Carlyle Moore

10	Modeling of Burning Particles in the Plume of a Flare 210-11MG-122	Dr. Olin Norton
11	Image Sequence Noise Reduction using Morphological Filters 210-11MG-105	Dr. Richard Peters
12	Laser-Induced Fluorescence of Hydroxyl 210-11MG-039	Dr. Chun Fu Su

Engineering and Services Center

13	Non-Volatile Precursors to Olefinic Chlorofluorocarbons [NVP-OCFCs] as Alternative Fire Extinguishing Agents with Reduced Global Environmental Impacts 210-11MG-113	Dr. William Bannister
14	Synthesis of Potential Intermediates of Microbial Biodegradation Pathways of Nitroaromatic Compounds 210-11MG-094	Dr. Joseph Dreisbach
15	Propanotrophic Biodegradation of Trichloroethylene (TCE) in Continuous Flow Reactors 210-11MG-067	Dr. Michael McFarland
16	Submicron Antennas for Solar Energy Conversion 210-11MG-045	Dr. Perry McNeill
17	The Microstructure of Compacted Moist Sand and its Effect on Stress Transmission 210-11MG-073	Dr. George Veyera

Frank J. Seiler Research Laboratory

18	Photoacoustic Calorimetry, ESR, and DSC Studies of NTO 210-11MG-055	Dr. Theodore Burkey

19	Chemical Reactivity and Thermodynamic Stability of Alkali Metals Deposited from Chloroaluminate Ambient-Temperature Molten Salt Electrolytes 210-11MG-034	Dr. Richard Carlin
20	Issues in the Application and Design of Reaction Mass Actuators 210-11MG-111	Dr. Ephrahim Garcia
21	A Control Formulation for the Active and Passive Damping of Flexible Structures 210-11MG-117	Dr. Thomas Posbergh

Volume II
Phillips Laboratory
Geophysics Directorate

22	PBL Disturbances over the Desert Southwest 210-11MG-060	Dr. Chia-Bo Chang
23	Effect of Magnus Moments on Missile Aerodynamic Performance 210-11MG-126	Dr. George Jumper
24	Simulation of the Spectra of Diatomic Ions and Molecules 210-11MG-118	Dr. Charles Lishawa
25	Plasma Transport in the Equatorial Ionosphere during the Great Magnetic Storm of March, 1989 210-11MG-109	Dr. Craig Rasmussen
26	Resonance-Enhanced Multiphoton Ionization of Atmospheric Molecules 210-11MG-038	Dr. Glenn Stark
27	Non Uniform Clouds 210-11MG-027	Dr. John Wills

Rocket Propulsion Directorate

28	Molecular Dynamics Study of Rocket Propellant Adhesion 210-11MG-032	Dr. Shannon Lieb

29	Integration of Prototype Reaction Wheel Torquers into the Astrex Facility 210-11MG-116	Dr. Thomas Pollock
30	Control System Design and Optimal Selection of Noisy Actuators and Sensors for Flexible Structures 210-11MG-089	Dr. Hung Vu
31	Sensor/Actuator Placement for Flexible Spacecraft 210-11MG-124	Dr. Trevor Williams

Advanced Weapons Survivability Directorate,
Lasers and Imaging Directorate, and
Space and Missle Technology Directorate

32	Second-Harmonic Generation in Corona-Poled Polymer Films 210-11MG-007	Dr. Gene Carlisle
33	Localization Limiters for the Microplane Concrete Material Model 210-11MG-035	Dr. William Cofer
34	No Report Submitted 210-11MG-004	Mr. Michael Houts (GSRP)
35	Misalignment Parameters of a Double Aperture Telescope Obtained as a Function of the Ratio of Optical Transfer Functions Without and With Diversity 210-11MG-133	Dr. Johanna Schruben
36	Applications of Compact Toroid Plasmas 210-11MG-082	Dr. Peter Walsh

Volume III
Wright Laboratory
Armament Directorate

37	Finite Element Analysis of Runway Failure Due to Blast Loading 210-11MG-110	Dr. Charles Camp
38	Analysis of the Penetration of Reinforced Concrete Slabs 210-11MG-125	Dr. Arnold Carden

39	Report Not Publishable at this Time 210-11MG-036	Dr. Charles Fosha
40	Methods which Accelerate Convergence in Iterative CFD Solvers 210-11MG-074	Dr. John George
41	Final Report for the Research Initiation Program 210-11MG-083	Mr. Randy Gove (GSRP)
42	Neural Networks for Guidance, Navigation, and Control of Exoatmospheric Interceptors 210-11MG-046	Dr. Kevin Moore
43	Fire Control System for a Laser Aimed Machine Gun 210-11MG-075	Dr. William Siuru
44	Robust Eigenstructure Assignment with Application to Missile Control 210-11MG-026	Dr. Kenneth Sobel

Aero Propulsion and Power Directorate

45	Evaluation of MOS-Controlled Thyristor (MCT) at 270 V DC for Static and Dynamic Loads 210-11MG-086	Dr. Muhammad Choudhry
46	Non Newtonian Effects of Powder Lubricant Slurries in Hydrostatic and Hydrodynamic Bearings 210-11MG-014	Dr. Donald Dareing
47	Investigation of the Combustion Characteristics of Confined Coannular Swirling Jets with a Sudden Expansion 210-11MG-013	Dr. Paul Hedman
48	Aircraft HVDC Power System - Stability Analysis 210-11MG-015	Dr. K. Sankara Rao

49	Dynamic Temperature Measurements in Reacting Flows 210-11MG-001	Dr. Larry Roe
50	No Report Submitted 210-11MG-098	Dr. Kaveh Tagavi
51	Droplet Distributions from the Breakup of a Cylindrical Liquid Jet 210-11MG-033	Dr. Richard Tankin
Avionics Directorate		
52	Function Decomposition and Machine-Learning Systems 210-11MG-041	Dr. Michael Breen
53	A Formal Process Model for Software Re-engineering: The Analysis Phase 210-11MG-068	Mr. Eric Byrne (GSRP)
54	Probability Modeling in Automatic Target Recognition 210-11MG-131	Dr. R. H. Cofer
55	The Enhancement of Connectionist Methods for Recognizing Airplanes from Radar Returns 210-11MG-006	Dr. Lawrence Hall
56	Report Not Publishable at this Time 210-11MG-135	Dr. Kevin Kirby
57	Fiber Laser Preamplifier for Laser Radar Detectors 210-11MG-134	Dr. Richard Miers
Electronic Technology Laboratory		
58	Computer Simulation of Small-Geometry High-Speed High-Density Integrated Circuit Performance Indicators 210-11MG-003	Dr. Ashok K. Goel
Flight Dynamics Directorate		
59	Studies on the Improvement of the Accuracy of Acoustic Emission Source Location for Smart Structures Applications 210-11MG-077	Dr. Marvin Hamstad

60	H-Infinity Control Design--A New Approach 210-11MG-017	Dr. Chin Hsu
61	Stress Wave Propagation Through the Thickness of Graphite/Epoxy Laminated Plates using PVDF Sensors 210-11MG-090	Dr. David Hui
62	Accelerated Fatigue Test Procedure for the Structural Polycarbonate Component of the F-16 Canopy Composite Material 210-11MG-005	Dr. Yulian Kin
63	Fatigue Fracture Behavior of Cord-Reinforced Rubber Composites 210-11MG-088	Dr. Byung-Lip Lee
64	The Efficacy of Constrained-Layer Damping Treatment to Suppress Parametric and Autoparametric Resonances in Nonlinear and Internally Resonant Nonlinear Structures 210-11MG-097	Dr. Lawrence Zavodney

Materials Directorate

65	AM1 Molecular Orbital Calculations of the Conformational Properties of Odd-electron Rigid Rod Polymer Model Species 210-11MG-002	Dr. John Connolly
66	Structural Analysis of Potential NLO Chromaphores 210-11MG-095	Dr. David Grossie
67	Eddy Current and Dielectric Spectroscopy Characterization of Metals and Ceramics - Application to NDE 210-11MG-085	Dr. Prasad Kadaba

68	Silicon/Tin Polymers for Enhanced Third Order Nonlinear Optical Properties 210-11MG-028	Dr. Joseph Lambert
69	Conformational Structure and Dynamics of Perfluoropolyalkylether Lubricants 210-11MG-020	Dr. Martin Schwartz
70	Modeling of the Formation of Macrosegregation During Casting Solidification 210-11MG-057	Dr. Hai-Lung Tsai

Volume IV
Aerospace Medicine Directorate

71	Empirical Prediction of Physiological Response to Prolonged Work in Encapsulating Protective Clothing 210-11MG-048	Dr. Phillip Bishop
72	Growth Dynamics of RAW 264.7 Mouse Macrophage Cells 210-11MG-076	Dr. Robert Blystone
73	Visual-Field Differences in Perception and Attention 210-11MG-022	Dr. Bruno Breitmeyer
74	Report Not Publishable at this Time 210-11MG-029	Dr. Vito DelVecchio
75	An Investigation of Dioxin Half-life Estimation in Veterans of Project Ranch Hand 210-11MG-065	Dr. Pushpa Gupta
76	Repeated Measures Design with Missing Observations 210-11MG-058	Dr. Ramesh Gupta
77	Pulmonary Effects of Hyperbaric Oxygenation 210-11MG-008	Dr. John Szarek

78	The Four-Artery Occlusion Model of Loss of Consciousness: Effects of Single and Multiple Episodes of Transient Cerebral Ischemia on Circle of Willis Blood Pressure, Neurochemistry and Behavior 210-11MG-108	Dr. Steven Waller
Crew Systems Directorate		
79	Sensitivity of Metabolic, Respiratory, and Performance Measures to Cognitive Demands in Dual-Task Performance 210-11MG-091	Dr. Richard Backs
80	Analysis and Modeling of Pilot Tasks and Decisions for the Pilot's Associate Pilot-Vehicle Interface using Hypermedia and a Constraints Based Approach 210-11MG-072	Mr. John Duncan
81	Effects of Delays in Networked Flight Simulators 210-11MG-114	Dr. Martin Hagan
82	Speaker Normalization using Neural Networks 210-11MG-019	Dr. Ashok Krishnamurthy
83	Modelling the Effects of Additional Head Mass on Head/Neck Dynamics 210-11MG-049	Dr. Amit Patra
Human Resources Directorate		
84	An Intelligent Tutoring System to Encourage Qualitative Reasoning in Basic Writing Students 210-11MG-040	Dr. Margaret Batschelet
85	Training a Complex Spatial Skill and Target Identification in both Single-Task and Dual-Task Modes 210-11MG-016	Dr. James Dykes
86	Computer Mediated Intellectual Teamwork 210-11MG-044	Dr. Verlin Hinsz
87	Job Survey Research Productivity Tool 210-11MG-051	Dr. DeLayne Hudspeth

88	Visually Significant Intraocular Auto Florescence 210-11MG-119	Dr. Gillray Kandel
89	Development of Key Variables for Multiship Simulation Benefit/Cost Analysis 210-11MG-100	Dr. William Moor
90	Shared Cognitive Representations of Teams 210-11MG-090	Dr. Joan Rentsch
91	Using Survival Analysis to Help Make Training Decisions 210-11MG-009	Dr. Stanley Stephenson

Occupational and Environmental Health Directorate

92	Mathematical Modeling and Decision-Making for Air Force Contaminant Migration Problems 210-11MG-084	Dr. Miguel Medina

FINITE ELEMENT ANALYSIS OF RUNWAY FAILURE DUE TO BLAST LOADING

Charles V. Camp
Principal Investigator

MEMPHIS STATE UNIVERSITY
Herff College of Engineering
Department of Civil Engineering
Memphis, TN 38152

December 20, 1991

INTRODUCTION

Accurate analysis and prediction of damage to runway pavements due to blast loadings from buried explosives is a complicated and difficult problem. A reliable model capable of accurately evaluating runway damage has to account for three distinct but interacting phenomena. First, during detonation of the explosive, suitable estimates of physical characteristics of the blast, such as the magnitude, the velocity, and the shape of the initial pressure wave, need to be determined. In the second phase of the model, the characteristics of the pressure wave are applied to the underlying pavement subgrade. As the blast wave spreads through the soil, the subgrade transmits a pressure to the underside of the runway pavement. In the final phase of the model, the transmitted pressure is applied as a dynamic distributed load on the concrete pavement. The structural response of the pavement is the most definitive mechanism in damage predictions.

The blast pressure loading is of a highly transient impulsive nature and produces a complex dynamic response from the concrete pavement. A level of damage for the runway system may be assessed from results of a nonlinear dynamic structural analysis of the concrete pavement. Although not listed as one of the three mechanisms in the damage model, the dynamic interaction between the soil subgrade and the structure is a critical component in the dynamic response of the pavement and probably the least understood.

Many researchers have developed simple single-degree-of-freedom (SDOF) and multiple-degree-of-freedom (MDOF) models for predicting the response of reinforced concrete structures to dynamic loads. Application of such models for predicting the damage of unreinforced concrete runways due to blast effects has had limited success. In most cases, the approximations used in formulating the simple model contribute to inaccurate results and an unrealistic structural response.

To develop a more reasonable model for the behavior of a runway subjected to a blast loading, the basic equations of motion for a continuum should be considered. Such an approach provides a fundamental characterization of the physical nature of the structure and how it responds to external loads. Constitutive relationships which describe the response of a material for a range of physical effects are coupled with the equations of motion to form a complete response mechanism.

The resulting formulation may be very complex and generally has few direct solutions. However, during the past few decades, researchers and engineers have used powerful numerical tools to solve these equations for a wide variety of problems in virtually all fields of science. The most robust of these numerical devices is the finite element method (FEM).

In this study, the primary objective is to review available FEM programs in an attempt to identify systems which are most appropriate for modelling the dynamic response of the concrete runways subjected to blast loading.

RESEARCH METHODOLOGY

The objectives of this project are twofold: first, to investigate the applicability of FEM systems to predict the damage of runway pavements due to buried explosives, and second, to recommend areas where future research will have the most significant impact. Upon completion of the review, a selected FEM program determined to be appropriate for analysis of runway pavements to dynamic loads is obtained and used to solve a simplified concrete pavement problem. From this analysis some measure of the operational performance may be ascertained. The project is divided into two phases:

Phase One

An survey of available FEM analysis programs is conducted to identify formulations which are capable of modelling problems similar to the runway damage problem. Information is gathered on each analysis package from the primary vendor. In addition, to vendor supplied information several experts in the fields of FEM modelling and the behavior of concrete materials are consulted.

The authors and/or distributor of the analysis software provided basic information on the fundamental science incorporated in their product. For example, the level of approximation of the governing equations and the type of material constitutive relationships used in the formulation. In particular, each analysis program is examined to determine how the dynamic response and failure of concrete is modelled.

Another important criterion used for this evaluation is the accessibility of basic code to modifications. Such a feature is advantageous and efficient for developing advanced material models. In runway damage predictions, the ability to incorporate and investigate different types of formulations to model the soil/structure interaction problem is an important consideration. Since reliable runway damage predictions must accurately model complex impulsive elasto-plastic material responses, fracture mechanics, and stress wave propagation, insights into how these mechanisms are treated is very important.

The ADINA System – Based on information supplied by the vendor and conversations with Dr. J. Tedesco, Professor of Civil Engineering at Auburn University, Auburn, Alabama, Dr. J. Walczak of ADINA R&D, Inc., and Dr. K. Bathe, Professor of Mechanical Engineering at the Massachusetts Institute of Technology, Cambridge, Massachusetts, the ADINA computer program system is selected as the most appropriate finite element analysis tool for preliminary runway damage predictions. The ADINA system has one of the most sophisticated concrete material models available and has the ability to analyze high–strain rate phenomena. Another important feature of the ADINA pro-

gram is the capability to incorporate user supplied loading and material modules into the general solution procedure. The ability to develop supplemental or new material models and incorporate these modules into the ADINA system is a critical feature for advanced studies of the runway damage problem.

ADINA Concrete Model – The general ADINA concrete material model may be used in both two- and three-dimensional analysis of structural systems characterized by large displacements and rotations. The basic concrete behavior model is described by several characteristics: a nonlinear stress–strain relationship, tension and compression failure envelops, and a model to control post-cracking and crushing of concrete.

In the ADINA concrete model, the general multiaxial stress–strain relationship is developed from a uniaxial stress–strain behavior. A typical uniaxial stress-strain relationship may be defined in terms of three strain phases: strain at ultimate tension cutoff, minimum crushing strain, and ultimate compressive strain. In a multiaxial stress condition, the stress–strain relationship is different depending on whether the material is being loaded of unloaded. In the complex material failure model for compressive failure, the form of the failure envelop is very sensitive to many user defined parameters. In most cases, the failure model parameters must be obtained from experimental data for the material and type of structure under consideration.

ADINA Distributed Loading – The ADINA system has modules to represent loading for concentrated, pressure, centrifugal and mass proportional, and user–supplied loading situations. The wide range of available loading configurations allows flexibility in selecting the most accurate way to represent the blast load on the concrete runway. The three-dimensional and the user defined loading modules are examined in the second phase of the this study. In large displacement analysis, the pressure loading may be a function of the deformation of the the structure. From observations, it is apparent that at some time after detonation, the runway slab separates from the underlying subgrade. At this point, the pressure loading is suspended. In addition, some form of centrifugal or mass proportional loading arising from the acceleration of the slab may be important in the analysis of the runway damage after the structure has lost contact with the subgrade.

Phase Two

The second phase of this project will be completed in the Summer of 1992. There are several reasons for requesting a six-month extension of the project. First and foremost was the difficulty Memphis State University and ADINA R&D, Inc. had in negotiating a special license of the ADINA system for University and education use. After several months of a settlement was reached and the ADINA software was delivered in September 1991. However, due to technical problems with the SUN SPARCstation version of ADINA final installation of the system was not completed until early December 1991. Therefore, phase two of the project will begin early January 1992.

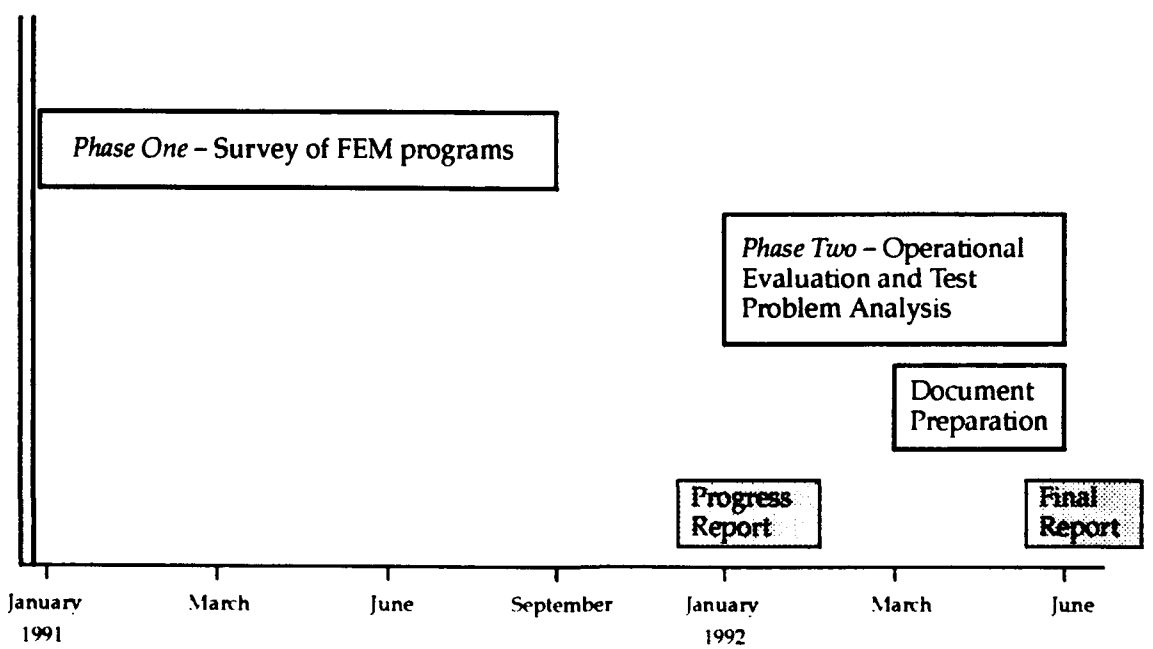
To establish some measure of the operational character of the ADINA system a simplified test problem will be analyzed. To maintain correlation with the actual runway problem, the dynamic response of a simply supported circular concrete plate subjected to a step concentrated load will be modelled. Several cases will be modeled. In the first case, the value of the step load will be less than the static pressure that will cause the plate to fail. In successive cases, the value of the unit load will be increased until the concrete plate fails. This series of test problems will provide common benchmarks for the evaluation of the performance and the operational properties of the ADINA system.

An important aspect in the operational evaluation is the practicality of pre- and post-processing graphics user interfaces. Well designed and written pre- and post-processing routines can greatly enhance the utility of large powerful numerical models. The ADINA system pre-and the post-processing routines will be assessed on their level of graphical and functional sophistication. As a primitive measure of programs portability, a list of the minimum computer configuration capable of executing an analysis using the ADINA system will be listed.

ANTICIPATED OUTCOMES - FUTURE DEVELOPMENTS

The results of this study will provide a source of important information on dynamic modelling of concrete pavements and directly influence the direction of future research efforts in predicting runway damage. Accurate assessment of the structural response of concrete pavements will provide an important tool for researchers and developers interested in evaluating the performance of runway designs.

TIME SCHEDULE (REVISED, 12/91)



RESEARCH BACKGROUND

This project is coordinated by the Herff College of Engineering at Memphis State University with the faculty of the Department of Civil Engineering conducting the investigation. The principal investigator for this project is Dr. Charles V. Camp, Assistant Professor of Civil Engineering. Professor Camp was a 1990 USAF Summer Faculty Research Fellow at Eglin AFB where he was involved in the analysis of runway damage using a simple model approach.

The Department of Civil Engineering is in the process of building a powerful network of workstation computers for the Structural Mechanics section. This facility is devoted to the analysis of structural systems and computational mechanics. Currently the structural mechanics computer laboratory consists of an upgraded SUN IPC workstation, a SUN SPARCstation 1+, and a SUN SPARCstation 2.

The research experience and capabilities of the MSU Civil Engineering faculty combined with the contributions of graduate and undergraduate students are sufficient to initiate and complete the proposed research within the project time period. The staff and resources of the MSU Civil Engineering Department will provide additional support in preparing the intermediate reports and the final project documents.

FINAL REPORT

TO

U E S, Inc.

FOR

1990 RESEARCH INITIATION PROGRAM

UNDER

U. S. Department of Defence - Air Force

(Through Universal Energy Systems)

ANALYSIS OF THE PENETRATION OF REINFORCED CONCRETE SLABS

Period of Performance: January 1, 1991 - December 31, 1991

#F49620-88-C-0053/SB5881-0378 P. O. #S-210--11MG-125

By

A. Eugene Carden

Department of Engineering Mechanics

The University of Alabama

Tuscaloosa, Al 35487-0278

ABSTRACT

ANALYSIS OF THE PENETRATION OF REINFORCED CONCRETE SLABS

The work performed under this grant was divided into three sections: Literature Survey, Study of the static and dynamic properties of concrete, and the use of the hydrocode, HULL, to perform a computer analysis of the penetration of reinforced concrete.

The literature survey consisted of three months work by Mr. Wanstall searching the Government Documents section of our library. Pertinent references are included with abstracts in the Final Report. There is some good work ongoing at several research institutions and there is need for additional work.

Mr. J. G. Lee has selected as his dissertation research topic the modification of the dynamic properties of concrete after exposure to moderate hydrostatic compression. An abstract has been submitted to the 1992 HVIS Symposium for consideration of the inclusion of a paper in the conference. Mr. Lee has developed a method for exposing concrete cylinders to about 8 kb hydrostatic pressure and then measuring the passage of an elastic wave through the material. The modifications in acoustic impedance and wave speed are then measured

We have modeled the reinforced concrete slabs in a 2D eulerian Hull analysis. In order to model the reinforcement, we have placed an equivalent amount of steel around the concrete slab to provide confinement. We ran one case with steel layers on entrance and exit side. During penetration the exit side steel detached from the concrete. We then modeled the reinforcement as a continuous shell around the slab, 48 cm thick and 98 cm in outer radius. This solved the detachment problem. We also provided a hole in the entrance side steel so that no work is done on initial penetration of that layer. We left no radial gap between the penetrator and entrance layer steel.

We do not obtain results in our computer analysis that match what we were shown on the test range at Eglin AFB. We do not see in the computer analysis significant bulging of the entrance side steel layer. The exit side steel layer does bulge but the geometries do not match the field data. We followed the event all the way to failure of the steel exit layer. Nor do we see the pressure pulse in front of the penetrator rise after the initial impact pressure begins to decay. Our analyses shows a persistent, monotonic drop in pressure that goes to zero. Velocities of 300 and 400 m/s have been tried and time periods of up to 5000 microsec have been followed. Furthermore, the failure model of concrete that we have used does not show the fracture locations that we have found documented in the literature. This work points the direction to further research.

FINAL REPORT

ANALYSIS OF THE PENETRATION OF REINFORCED SLABS

INTRODUCTION

This is a final report showing the results of our work on this project. Professor Carden and Mr. Wanstall spent a summer at Eglin AFB, 1990 performing work on this same topic. Much of the summer was spent learning a new computer system, Hull Hydrocode, and begining to learn how to model the problem in 2D eulerian format. We were taken to a test range and shown a large number of reinforced concrete slabs that had been penetrated by fairly low velocity projectiles.

The steel reinforcing mesh of the failed slabs seen on the range was bowed out on both entrance and exit side. It was our understanding of the problem that at low impact velocities, the shock pressures would be low, but the acoustic wave reflections combined with the volume change in the concrete (due to the displacement of the concrete by the presence of a solid rod) accounted for the build up of pressure in the bulk of the concrete in the vicinity of the penetrator. This pressure exerted itself against the steel reinforcement and pushed the reinforcement outward, both on the entrance and exit sides. The reinforcement was bent outward like a membrane under internal prressure. This led to an estimate of the pressure. But the build up of pressure in the concrete also builds up pressure on the front of the penetrator. At a velocity of 300 m/s, the time for penetration of a one meter thick slab is of the order of 5 to 15 ms. It is this build up of pressure that extracts momentum from the penetrator. Without the steel reinforcement, such a build up of pressure would not be possible since the pressure on the free surfaces of the concrete is essentially zero. We see the effect of the reinforcing rods (if spaced sufficiently close together in a 3D mesh) as a confinement for the concrete. Such mesh, for maximum benefit, must be on both entrance and exit sides of the slab. Consequently, long run times for the hydrocode are necessary to observe the penetration event. Long run times and small cell size require large computing facilities and patience.

We have performed these hydrocode analyses on an IBM RS 6000 with 32 mb of core memory and two hard disks, one of 360 mb and another of 1.2 Gb. We have used OTIHULL version 4 for the analysis.

Object and Scope. The purpose of this work was to perform hydrocode analyses of thick reinforced concrete slabs using low velocity projectiles. In the overall plan, a literature on concrete dynamic and static mechanical properties and techniques for measuring these properties were reviewed. This literature survey suggested that we needed to perform some dynamic property tests and we have built and calibrated a facility to to that.

LITERATURE SURVEY.

The results of the survey of the NTIS unclassified data base of concrete properties and test methods is included in an appendix. We found a number of articles of which we were unaware and a few that we deem as significant. The work by Rajendran (Dayton Univ.) and Nicolas (WPAFB) has some interesting innovations for formulating dynamic constitutive equations. The work by Malvern shows results from a large Split Hopkinson Pressure bar experiments on large samples of concrete. From these SHPB tests stress-strain functions for plane strain compression can be obtained at strain rates of up to about 2000 per sec. In general, the SHPB does not give information concerning fracture and the results are limited to strains of up to about 10 percent. No information concerning failure under shearing stresses (Mode II, III fracture) was found. Little information concerning the heterogeneity of concrete and its influence on wave propagation and attenuation was found. It is known that ordinary concrete has a significant pore volume. Pressures above some threshold will begin to close these pores and raise the density. No references could be found that elucidate this fact and its significance to dynamic loading. The only published Hugoniot on concrete was that produced at General Motors Research Institute several decades ago.

DEVELOPMENT OF A DYNAMIC TEST METHOD FOR CONCRETE

We have seen the need to measure and record the dynamic properties of concrete for elastic waves. If one models the low velocity impact of thick reinforced concrete slabs, one observes that the elastic wave reflects between the reinforcement planes about twenty times. While spherical expansion attenuates the particle velocity and pressure, the reflection from a high acoustic impedance surface (the steel reinforcement plane) causes an increase in the pressure. Furthermore, the displacement of the concrete caused by the insertion of the projectile causes an increase in pressure. For unconfined concrete, modeled as a cylinder under pressure, the maximum pressure is governed by the tensile strength (static pressure) or the maximum shearing strength (for dynamic loading). The equation of motion is:

$$(d\sigma_r/dr) = (\sigma_\theta - \sigma_r)/r + \rho \ddot{r}$$

The $(\sigma_\theta - \sigma_r)$ term is twice the maximum shearing stress for the state of stress. For dynamic loading of a cylinder, with significant acceleration terms, both stresses would be negative on the inside surface of a cylinder. The hoop stress will be tensile on the outside surface. This was pointed out by G. W. Taylor in an article on the fracturing of bombs. Twice tau max is the difference beween the two principal stresses. For the metals, fracture begins on the outside surface. We address this model only to point out that there is a limit to the pressure gradient in concrete determined by the geometry, the boundary conditions, the acceleration terms and the strength. For low velocity impact of UN reinforced concrete, this pressure gradient

is quite low. But for reinforced concrete, the boundary value of pressure is whatever is limited by the confinement properties of the reinforcement. Therefore, the reinforcement allows a much greater pressure region in front of the penetrator after some initial period of time to allow this pressure to build up. The time depends upon the dimensions of the slab, the size of the penetrator, and the geometry of the reinforcement.

We determined that a clearer understanding was needed for the elastic propagation (and reflection) of impact waves in concrete. Especially as affected by the changes wrought in the concrete by the pressure excursions. We have therefore modified a drop tower to impact a 2 in dia concrete cylinder mounted on a steel load bar and impacted from the top by a two inch dia steel bar having velocities of up to 10 m/s. Because of the difference of steel and concrete, the wave passing through the concrete reflects at the next concrete-steel interface. It reflects back and forth until the transmitted wave in the load bar reaches a stabilized value. The intensity of the first reflection is a measure of the impedance mismatch, and the time period between the reflections is a measure of the propagation velocity. Prior exposure to high hydrostatic pressure causes a permanent change in density. The compressive elastic modulus also increases due to the increase in density. If these do not change proportionally, there is a change in the propagation velocity, and there will be an increase in acoustic impedance. A dissertation by Mr. J. G. Lee is to be published which will show these methods and data. We show in Appendix some of the data taken for a steel on a steel impact where the steel specimen has an area mismatch, not an impedance mismatch.

HYDROCODE ANALYSIS OF THE IMPACT OF REINFORCED CONCRETE

We have modeled the concrete slab as a 2D eulerian cylinder, 48 cm thick, 96 cm in radius. The penetrator was modeled as a rigid island 4 cm in radius and 50 cm long. At Eglin AFB we ran several analyses out to 250 microsec. We never achieved more than about 20 percent penetration. It became apparent that we should continue the problem until complete penetration was achieved. Furthermore we could not model a 3D reinforcement mesh in a 2D calculation. We therefore calculated the "equivalent" thickness of steel represented as a continuous sheet. At first the steel was only on the exit side. But this sheet separated from the concrete when the reflected stress in the concrete was tensile. We solved this by making the steel continuous around the whole concrete slab. Next we opened a hole in the entrance side steel sheet so that the penetration of the entrance side steel sheet did not influence the problem significantly. We have several significant observations.

1. The mesh size affects the results. We tried 0.1 cm cell size, but that gives us 1 E6 number of cells and our machine will not run the problem incore. We later tried mixed cells, but at the interfaces where the cell sizes change (0.2 to 0.3 to 0.4 cm) we see discontinuities in pressure and density.

2. The station data show a monotonic decrease in pressure after the initial pressure pulse (after about 30 microsec). Even

though the penetrator bulges the steel and produces a fracture (at about 5000 microsec) the pressure in front of the penetrator does not rise. We are not sure why this is so.

3. The steel sheet on the entrance side shows a very small amount of "bulging". We believe that the displaced volume of the concrete caused by the insertion of an 8 cm dia x 50 cm long penetrator, should cause the pressures to rise, just as when the penetrator approaches the exit side steel sheet we believe that the pressure should rise. Our output shows the pressures are inordinately low.

4. Our current material library on concrete does not contain a fracture model and the constitutive equation does not give realistic results.

5. The steel sheet on the exit side show bulging and fracture but the region of distortion is limited to about 3 diameters of the penetrator. In the field observations of failed slabs, the distortion of the reinforcing steel extended at least ten diameters of the penetrator.

6. In our computer analyses the concrete flows past the penetrator and does not scrape or join the lateral surfaces. This is consistent with the observation of recovered penetrators.

We include some of the density contours and station data for several of our runs.

SUGGESTIONS FOR FUTURE WORK

The results of this work suggest that more work needs to be performed to find an analysis that conforms to the observations of field tests. We suggest the following:

All of our runs have been with an eulerian mesh. We have not evaluated the differences that would result from using a lagrangian mesh. We know that cell size has a significant influence on the results. We suggest a 0.1 cm cell size and an out of core calculation. This will greatly increase the time for the calculation. Furthermore, some defaults in the Hull main program have to be changed to get the program to continue after long run times.

We have obtained another material library for concrete. We suggest that the results of two identical runs be compared with these two sets of materials data.

We also suggest that the Hull program be used to evaluate low velocity impact of steel - concrete - steel cylinders as we are currently using in our experimental work. It is not known whether any hydrocode will propertly treat elastic impact.

We suggest that a dynamic constitutive equation be developed. This function must be history dependent. Increases in density are partly recoverable and partly permanent. Evidently the work at the University of Dayton is addressing some of these questions.

ACKNOWLEDGMENTS

Mr. Wanstall performed nearly all of the literature survey and the initial set up of the Hull program on the RS 6000. Mr. J. G. Lee is performing the dynamic tests of concrete for a dissertation. He has also performed all of the Hull runs during the last two months of the grant period. Dr. J. L. Hill, Head of Engineering Mechanics provided us with first priority on an RS 6000 machine so that we could perform these hydrocode analyses. Dr. Rob Smith assisted in getting the RS 6000 installed, the Hull program installed and running. Appreciation for the efforts of each of these is gratefully given.

APPENDICES

I. LITERATURE SURVEY REFERENCES AND ABSTRACTS

II. SAMPLE RESULTS FROM DYNAMIC EXPERIMENTS WITH CONCRETE

III. HULL OUTPUT ON SELECTED RUNS

Report # 39
210-11MG-036
Prof. Charles Fosha
Report Not Publishable at This Time

METHODS WHICH ACCELERATE CONVERGENCE IN ITERATIVE CFD SOLVERS

A FINAL REPORT TO THE AFOSR RIP PROGRAM

Author: John H. George
Professor of Mathematics
Department of Mathematics
Box 3036
University of Wyoming
Laramie of Wyoming 82071

ABSTRACT

Several solution methods have been investigated which yield an exact solution (to machine accuracy) of the backwards Euler implicit solution method of the Navier-Stokes equations. These methods include the generalized minimum residual (GMRES) [11, 19], the conjugate gradient (CG) [6], and the conjugate gradient squared (CGS) [14]. These methods have been coded, and applied to several versions of; first, a prototypical model, second, many versions of actual flow code. A version of the CGS method was found to be the most robust and gives the best convergence properties on the computational fluid dynamics (CFD) class of problems. The method was applied to Whitfields' numerically computed Jacobian, and compared well with conventional methods. We have applied the approximate derivative concept to the CGS solution method, and after the solution is well formed (after the time marched solution is very close to a steady state), the solution works well. There are problem in obtaining quadratic convergence, the expected convergence for Newton's method. During the early unformed solution, the Newton iteration does not converge with a full Newton step. This means relaxation parameters which limit the step size must be determined to optimize the convergence in this early phase.

If proper preconditioning were obtained, the convergence in the early steps should improve.

I. INTRODUCTION:

The numerical solution of the Navier-Stokes equations over complex multiple bodies [1,3,9], is usually based on a backwards Euler-Implicit solution method that proceeds in time (or pseudo- time) to a steady or unsteady state solution [15]. The final optimum solution does not currently exist, and will always need some user input to decide the combination of 'best' algorithms for a particular configuration and boundary conditions. Often, a steady state solution is sought of the motion around the particular configuration. In this process, it is not so important to obtain a time accurate solution. The numerical solution method involves several complex techniques which will now be described. 1. A transformation from physical space to computational space (rectangular) is introduced. This allows proper application of the correct boundary conditions to the body.
2. Solving the resulting system of equations. The class of methods that are currently most effective are backward Euler-Implicit schemes, utilizing some version of an LU splitting of the operator. The resulting nonlinear system is solved by Newton's method [4,5,7]. This research has added the (CGS) solution method to this class of solution methods, and can iteratively attain any amount of precision at each step in the solution process. This allows a much larger CFL number than that permitted by the approximate LU factorization (the divergence in the standard LU factorization occurs at about CFL=10, according to Barth [2]).

II. OBJECTIVES OF THIS RESEARCH EFFORT:

The problem of speeding the convergence of the flow solver is a complex problem due to the strong nonlinearities and the large dimensionality of this class of problems.

The main thrust of this research was to;

1. include the CGS method in a variety of CFD code developed by Dave Whitfield of Mississippi State and amoung others, his Air Force colleagues Dave Belk, Bruce Simpson, and Kirk Vanden [5,12,16].

2. apply the CGS solution method to a variety of realistic flow and boundary conditions, both for steady and unsteady state solutions.

3. obtain quadratic convergence using the conjugate gradient CGS algorithm. This has produced limited success, without any preconditioning, converging only after the solution is near to a steady state configuration. Much additional work is needed here to determine the limits of the direct application of Newtons' method to obtaining CFD steady-state solutions.

III. ORGANIZATION:

When using the conventional LU splitting and the resulting error, the final solution is not as accurate as it could be if an 'exact' iterative solver is used As was pointed out earlier, the numerical experimental limit on the CFL number using approximate LU factorization is about 10. The use of solvers such as the conjugate-gradient squared (CGS) method which require for solution only multiplication of the resulting Jacobian by an arbitrary vector is a promising alternative to LU solution methods and allows a large CFL number in most problems. Less complex versions of this technique, the CG method, has had excellent success in complex oil reservoir simulations, and in many other large scale simulations.

Any improvement in the computational efficiency of a typical flow code should have immediate impact on the entire class of CFD numerical methods. The inclusion of the CGS 'exact' solution method into a large variety of CFD code as provided by this research is a major step in this direction.

By solving the problems with more accuracy at each step, the resulting trek to steady state can be done with a higher degree of accuracy. The same technique yields more accurate solutions to unsteady problems (time accurate). The important advantage to using conjugate gradient class solvers is that the computations can be easily vectorized on CRAY style computers. As the preliminary part of this research was to establish that the CGS conjugate gradient methods were viable and produced useful results, the codes were not vectorized.

The application of these solution methods to CFD problems is complicated by the fact that the highly nonlinear interactions cause many problems [1]. The main class of methods we are looking at involve using the conjugate gradient method to accelerate existing code. The idea is that oscillations caused by complex eigenvalues can greatly slow the convergence history. As the problem complex eigenvalue propagates through the system of nonlinear equations, by minimizing the new residual using a fixed set of residuals from past iterations, The determination of the parameter k is critical as it determines the number of eigenvalues that will be 'filtered' out of the iteration. The generalized minimum residual concept is due to Saad and Schultz [11]. A nonlinear version of the generalized minimum residual, is being studied for inclusion in existing code [7]. The conjugate gradient method will theoretically produce a solution to the equation in a finite number of steps, while preserving the sparsity structure of the matrix. Thus, the conjugate gradient method is in a sense, an exact solution method.

IV RESULTS:

This work established that conjugate gradient methods, such as the CGS method are effective and can give 'exact' solutions to the Newton's iteration needed to solve the backwards Euler method.

Particular code that has been modified to run the CGS solver are;

the Redcoon and Juli codes which use common blocks to pass data.

several versions of the Ruby code that passes data in the subroutine arguments.

A time accurate version of Ruby which produces time accurate solutions at each iteration.

Several of these codes have been ported to 386 pc's with extended memory. The code were run on simple configurations and much computer time on the CRAY was saved. The preliminary testing was done on the smaller, slower pc's. The final versions were ran on the CRAY.

The use of approximate derivatives as generated by Whitfield were solved with no problem. When the approximate derivatives were used in the CGS to calculate the Jacobian approximately, we had trouble converging. Since we were using the origional preconditioner, instead of a numerically generated preconditioner as was done in Whitfields' version, it is felt that the difficulties can be overcome by designing a numerical preconditioner. A similar approach was considered in Obayashi [10].

All of these modified code are stored on the mass storage device on the Eglin CRAY and are available to the Air Force personel.

IV RECOMMENDATIONS:

While the application of the conjugate gradient class of solution CGS method to a real flow solver has produced results on every problem it was applied to, the full vectorization of the CGS method is still to be done. This will increase the speed of the calculations on the CRAY machine.

The study of the best preconditioners to use is only begining, and the more work done here, the less iterations will be needed. The use of preconditioners that do not need the Jacobian are being investigated and look promising.

The selection of proper preconditioners and more work on the calculation of the numerical derivatives in the CGS method will lead to quadratic convergence. This will require a special application of the solver, one which will require a closer connection to the standard Newton method which does have quadratic convergence.

The use of conjugate gradient solution methods is a promising step to furthering the advancement of computational fluid dynamics (CFD) simulations.

Domain decomposition methods are easily exploited by the CGS method. These concepts, (see [8]) should prove useful in CFD with complicated multi-body geometries.

The ideas of vector sequence convergence [13] are extremely important, and require further study.

REFERENCES:

[1] Anderson, D.A., Tannehill, J.C., and Pletcher, Richard H., Computational Fluid Mechanics and Heat Transfer, Hemisphere Publishing Corporation, New York, 1984.

[2] Barth, T.J., ''Analysis of implicit local linearization techniques for upwind and TVD algorithms '', AIAA-87-0595.

[3] Bunning, P., ''Computation of inviscid transonic flow using flux vector splitting in generalized coordinates'', Phd Thesis, Dept of Aero-Astro Stanford, October, 1983.

[4] Dennis, J. E., and Schnabel, R. B., ''Numerical Methods for Unconstrained Optimization and Nonlinear Equations,'' Prentice Hall Series in Computational Mathematics, Prentice Hall, Englewoods Cliff New Jersey, 1983.

[5] Gatlin, B. and Whitfield, D.L., ''An Implicit, Upwind, Finite-Volume Method for Solving The Three-Dimensional Thin-Layer Navier-Stokes Equations,'' American Institute of Aeronautics and Astronautics Paper AIAA-87-1149, June 1987.

[6] Golub, G.H., and Van Loan, C.F., Matrix Computations, 2nd ed. John Hopkins University Press, Baltimore Maryland, 1989.

[7] Huang, C-Y.,Kennon, S. R. and Dulikravich, G. S., ''Generalized nonlinear minimum residual (GNLMR) method for iterative algorithms'', J. Comp. and Appl. Math. 16 pp. 215-232 (1986).

[8] Meurant, G., ''Domain decomposition methods for partial differential equations on parallel computers'', Int. J. of Supercomput. Appl. 2 No. 4 pp. 5-12 (1988).

[9] MacCormack R.W.,''Current status of numerical solutions of the Navier-Stokes equations '', AIAA-85-0032.

[10] Obayashi, S., ''Numerical simmulation of underexpanded plumes using upwind algorithms'', AIAA-88-4360.

[11] Saad,Y., and Shultz, M. H.,''GMRES: A generalized minimum residual algorithm for solving nonsymmetric linear systems'', SIAM J. SCI. STAT. COMPUT. 7 pp. 856-869 (1986).

[12] Simpson, L.B. and Whitfield, D.L., ''A Flux Difference Split Algorithm for Unsteady Thin-Layer Navier-Stokes Solutions,'' AIAA-89-1995, 9th Computational Fluid Dynamics Conference, June 1989.

[13] Smith, D. A.,Ford, W. F. and Sidi, A.,''Extrapolation methods for vector sequence,'' SIAM Review 29 No. 2 pp. 199-233 (1987).

[14] Sonneveld, P.,''CGS, a fast Lanczos-type solver for nonsymmetric linear systems'',SIAM J. SCI. STAT. COMPUT. 10 pp. 36-52 (1989).

[15] Strikwerda, J.C.,Finite Difference Schemes and Partial Differential Equations, Wadsworth and Brooks/Cole Mathematics series, Pacific Grove California, 1989.

[16] Whitfield, D.L., ''Implicit Upwind Finite Volume Scheme for the Three-Dimensional Euler Equations,'' Mississippi State University Report, MSSU-EIRS-ASE-85-1, September 1985.

[17] Venkatakrishnan, V., ''Preconditioned conjugate gradient methods for the compressible Navier-Stokes Equations,'' AIAA 1990.

Final Report for Research Initiation Program

Contract No. S-210-11MG-083

From:
Department of Physics
The University of Alabama in Huntsville
Huntsville, AL 35899

Principal Investigator
Professor Russell A. Chipman

Investigator
Shih-Yau Lu

Dec 23, 1991

Executive Summary

The Executive Summary provides a brief synopsis of the content of the Final Report (Contract No. S-210-11MG-083).

This research was motivated by the need to appropriately interpret the Jones and Mueller matrices and develop algorithms to extract polarization properties from these matrices.

It is well known that a polarization element can be charactered by either a Jones matrix or a Mueller matrix. In fact, in many laboratories, the Mueller matrices are now being routinely measured. However the polarization properties described by a Jones or Mueller matrix are not apparent. The interpretation of the Jones and Mueller matrices then becomes crucial, but not yet clear. Nevertheless, two useful quantities- diattenuation and retardance, were introduced for this purpose by several authors. They are used to quantify, respectively, the amplitude and phase properties of the homogeneous polarization elements. However, the definition and meaning of these quantities are not evident for inhomogeneous polarization elements. The inhomogeneous polarization elements denote the polarization elements having nonorthogonal eigenpolarizations. The importance of inhomogeneous polarization elements can not be overemphasized. First, combinations of polarization elements are, in general, inhomogeneous. Also, skew rays in optical systems tend to be slightly inhomogeneous. Understanding inhomogeneous polarization elements is important in understanding depolarizing elements, since it is necessary to separate the inhomogeneous part of the Mueller matrix from depolarizing part in data reduction for depolarizing elements.

To interpret Jones and nondepolarizing Mueller matrices, particularly those describing inhomogeneous polarization elements, is the aim of this research. The outline of this research consists of analysis of the exponential representation of Jones

matrices, analysis of inhomogeneous polarization elements, and the data reduction of Jones and Mueller matrices.

The following is an outline of the work contained in this final report. When the Jones matrix is expressed in exponential from $J = \exp(d_0\sigma_0 + d_1\sigma_1 + d_2\sigma_2 + d_3\sigma_3)$, then the real part and imaginary part of the d-coefficient have simple physical meaning. The real parts of d-coefficients correspond to diattenuation and imaginary parts correspond to retardance. By applying Baker-Campell-Hausdorff identity, it is shown that this representation has several interesting order independent properties and is very useful in addressing deeper questions.

From the polar decomposition theorem of matrix theory, any polarization element, homogeneous or inhomogeneous, is equivalent to a cascade of diattenuator and retarder, and then the diattenuation and retardance of the polarization elements can be unambiguously defined. We introduce a new unitary invariant, called inhomogeneity, which measures the nonorthogonality of the eigenpolarizations. Explicit expressions for diattenuation and retardance have been derived in terms of the unitary invariants- the eigenvalues and the inhomogeneity. The resulting expressions are suitable for performing data reduction on Jones and nondepolarizing Mueller matrices. Finally it is shown that two Jones matrices which share the same eigenvalues and inhomogeneity are related by a unitary transformation, or in physical terms, one polarization element can be transformed into the other through the addition of a pair of orthogonal retarders, one on either side of the element. We term this the unitary equivalence theorem of Jones matrices. Based upon this theorem, the classification of Jones matrices by unitary invariants is established. The generalization of this results to include depolarizing elements is now in progressing, and some preliminary results have been found.

The achievements of this research are the following: First, several interesting prop-

erties of exponential representation of Jones matrices are found. Second, the definition of diattenuation and retardance are unambiguously generalized to include inhomogeneous polarization elements. Third, a measure of inhomogeneity is introduced. Finally a new classification scheme for Jones matrices is established.

The remainder of this final report consists three parts. First is a draft copy of our paper in preparation for the Journal of the Optical Society, "Generalized diattenuation and retardance for inhomogeneous polarization elements". Section two and three contain studies on the Jones and Mueller calculus which include the results on the exponential representation of the Jones matrix. As these studies are completed, further manuscripts will be prepared for refereed journals. The last section contains the materials from a presentation regarding this research given at Eglin Air Force Base in December, 1991.

Table of Contents

1. Generalized diattenuation and retardance for inhomogeneous polarization elements
2. Studies on the Jones calculus
3. Studies on the Mueller calculus
4. Presentation materials

Generalized diattenuation and retardance for inhomogeneous polarization elements

Shih-Yau Lu and Russell A. Chipman
Department of Physics
University of Alabama in Huntsville
Huntsville, AL 35899

ABSTRACT

Two useful quantities- diattenuation and retardance, that characterize the amplitude and phase properties of a polarization element, are generalized to inhomogeneous polarization elements, i.e., the polarization elements with nonorthogonal eigenpolarizations. Their explicit expressions are obtained, and they are suitable for performing data reduction on the Jones and nondepolarizing Mueller matrices. Besides, classification of polarization elements by their unitary invariants is discussed.

1. INTRODUCTION

In this paper the polarization element (PE) denotes an optical element or a cascade of optical elements that can modify the polarization state of incident light. In this paper we treat only nondepolarizing elements. PEs can be classified into two types, namely homogeneous and inhomogeneous[1,2], based upon the orthogonality of their eigenpolarizations. These terminologies were introduced by Shurcliff who used them to classify polarizers[2]. A homogeneous PE has two orthogonal eigenpolarizations. Most of polarizers and retarder are homogeneous. Homogeneous and isotopic interfaces, such as lens and mirrors, are homogeneous with linear eigenpolarizations. An inhomogeneous PE has nonorthogonal eigenpolarizations. In many cases, an inhomogeneous PE results from a cascade of several different homogeneous PEs. Also, skew ray paths through optical systems intend to be slightly inhomogeneous. An inhomogeneous PE, which has only one eigenpolarization, is called degenerate[3]. It is common to characterize a PE by either a Jones matrix or a Mueller matrix. The Jones and Mueller matrices explicitly show how incident Jones and Stokes vectors map into transmitted Jones and Stokes vectors. However the diattenuation and retardance described by a Jones or a Mueller matrix are not readily apparent. The interpretation of Jones and Mueller matrices then becomes crucial, but it is not yet clear. Thus it is highly desirable to define the diattenuation and retardance properties of a PE and express them explicitly in terms of the matrix elements of the Jones or Mueller matrix.

In general, a PE does not only modify the polarization state of the incident light, but, consequently, it also changes the intensity and phase of the incident light. Two useful quantities, diattenuation and retardance, that reflect the intensity and phase properties of

a homogeneous PE, are summarized in Ref. 1. According to Ref. 1, the diattenuation $\mathcal{D}$ and the retardance $\mathcal{R}$ of a homogeneous PE are respectively defined as

$$\mathcal{D} = \frac{\left||\xi_q|^2 - |\xi_r|^2\right|}{|\xi_q|^2 + |\xi_r|^2}, \quad 1 \geq \mathcal{D} \geq 0, \tag{1}$$

$$\mathcal{R} = |\delta_q - \delta_r|, \quad \pi \geq \mathcal{R} \geq 0, \tag{2}$$

where $\xi_q = |\xi_q| \exp(i\delta_q)$ and $\xi_r = |\xi_r| \exp(i\delta_r)$ denote the eigenvalues of the Jones matrix of this PE. Throughout this paper we use $\mathcal{D}$ and $\mathcal{R}$ to denote diattenuation and retardance. It should be noted that $|\xi_q|^2$ and $|\xi_r|^2$ are the maximum and minimum intensity transmittance of the homogeneous PE. Eqs. (1) and (2) imply that diattenuation is a measure of the dependence of the PE's intensity transmittance upon the incident polarization state, and retardance is a measure of the dependence of the PE's optical path length upon the incident polarization state. Besides, the diattenuation defined in Eq. (1) is equal to the degree of polarization (DOP) produced by the homogeneous PE when incident light is unpolarized[1]. Thus, diattenuation can be regarded as a generalization of the polarizance which was also introduced by Shurcliff[2] as a performance factor for a polarizer. The homogeneous PE of zero diattenuation is called a (homogeneous) retarder, and the homogeneous PE of zero retardance is called a (homogeneous) diattenuator. However, Eqs. (1) and (2) can not be applied to inhomogeneous PEs. It is the main purpose of this paper to generalize the definition of diattenuation and retardance to include inhomogeneous PEs, and develop algorithms to calculate them from these matrices.

The outline of the approach in this paper are the following. By applying the polar decomposition theorem of matrix theory, any PE, homogeneous or inhomogeneous, is

equivalent to a cascade of diattenuator and retarder, and then the diattenuation and retardance of the PE can be unambiguously defined. Moreover, we introduce a new unitary invariant called inhomogeneity which is a measure of the inhomogeneity. Geometrically it relates to the nonorthogonality of eigenpolarizations. The explicit expressions for diattenuation and retardance in terms of unitary invariants- eigenvalues and inhomogeneity, are obtained. Finally, the unitary equivalence theorem are readily formulated.

In fact, Gil and Bernabeu[4] also used polar decomposition theorem to define the polarizing and retardation parameters for nondepolarizing Mueller matrices. Though the approach in this paper is similar to theirs, results of this paper are more useful. First only two parameters, diattenuation and retardance, instead of seven are used to characterize the polarization properties of a PE. We also obtain the expressions for diattenuation and retardance in terms of the unitary invariants of the Jones matrix. In addition, a parameter that is an indicator of inhomogeneity is introduced. Compared with other classification schemes[3,4,5], the classification discused in this paper has the advantage that it is based upon the unitary invariants of Jones matrices.

2. DIATTENUATON AND RETARDANCE FOR AN INHOMOGENEOUS PE - An Application of the Polar Decomposition Theorem

From the polar decomposition theorem, any PE is equivalent to a cascade of diattenuator and retarder. Furthermore, the diattenuation and retardance of a PE can be, respectively, defined to be corresponding quantities of the diattenuator and retarder in polar decomposition.

Consider an arbitrary PE whose Jones matrix is denoted by $\mathbf{J}$. According to the polar decomposition theorem[4,6,7], the Jones matrix can be written as

$$\mathbf{J} = \mathbf{J}_R \mathbf{J}_D , \tag{3}$$

or

$$\mathbf{J} = \mathbf{J}_D' \mathbf{J}_R \tag{4}$$

where $\mathbf{J}_D$ and $\mathbf{J}_D'$ are nonnegative define Hermitian matrices and $\mathbf{J}_R$ is unitary. Moreover, $\mathbf{J}_D$ and $\mathbf{J}_D'$ are respectively given by[7]

$$\mathbf{J}_D = \sqrt{\mathbf{J}^\dagger \mathbf{J}}, \tag{5}$$

and

$$\mathbf{J}_D' = \sqrt{\mathbf{J}\mathbf{J}^\dagger} \tag{6}$$

where the dagger denotes the Hermitian conjugate. In general, $\mathbf{J}_D$ and $\mathbf{J}_D'$ are not identical, but their eigenvalues are identical. If the determinant of $\mathbf{J}$ is not zero, then $\mathbf{J}_R$ is uniquely given by

$$\mathbf{J}_R = \mathbf{J}\mathbf{J}_D^{-1} = \mathbf{J}_D'^{-1}\mathbf{J} \tag{7}$$

or, equivalently,

$$\mathbf{J}_R = \frac{\mathbf{J} + (\mathrm{Det}\,\mathbf{J})(\mathbf{J}^\dagger)^{-1}}{\mathrm{Tr}\,\mathbf{J}}, \tag{8}$$

where Det and Tr denote the determinant and trace, respectively. $\mathbf{J}_R$ is not unique when the determinant of $\mathbf{J}$ is zero, but it will be pointed later that ambiguity exits. Since $\mathbf{J}_D$ and $\mathbf{J}_D'$ correspond to diattenuators, and $\mathbf{J}_R$ corresponds to a retarder, any PE can be interpreted as a cascade of retarder and diattenuator.

Since a retarder does not change the DOP or intensity of incident light, diattenuation of the PE is equal to the diattenuation of $\mathbf{J}_D$ (or $\mathbf{J}_D'$). The intensity transmittance of the PE can be written as

$$T = (\mathbf{J}\hat{\mathbf{E}})^\dagger (\mathbf{J}\hat{\mathbf{E}}) = \hat{\mathbf{E}}^\dagger \mathbf{J}^\dagger \mathbf{J}\hat{\mathbf{E}}, \tag{9}$$

The Jones vector with karat is normalized. It can be realized that the eigenvalues of J_D (or $J_D^{\dagger}$) are the square root of the PE's maximum and minimum intensity transmittance, T_{max} and T_{min}. Therefore, the appropriate definition of the diattenuation for the PE is

$$\mathcal{D} = \frac{T_{max} - T_{min}}{T_{max} + T_{min}}, \quad 1 \geq \mathcal{D} \geq 0. \tag{10}$$

This definition is a generalization of Eq. (1), and Eq. (10) reduces to Eq. (1) for homogeneous PEs. Let $\hat{E}_{max}$ and $\hat{E}_{min}$ be the incident polarization states which yield the maximum and minimum intensity transmittance. Since they are the eigenpolarizations of J_D, it follows that

$$\hat{E}^{\dagger}_{max}\hat{E}_{min} = 0, \tag{11}$$

and

$$(J\hat{E}_{max})^{\dagger}(J\hat{E}_{min}) = 0. \tag{12}$$

Thus diattenuation defined in Eq. (10) is always equal to the DOP of the output light when the incident light of the PE is unpolarized.

Similarly, the retardance of the PE is defined to be the retardance of J_R, and it can be written as

$$\mathcal{R} = 2\cos^{-1}\frac{|Tr\, J_R|}{2}, \quad \pi \geq \mathcal{R} \geq 0. \tag{13}$$

in which J_R is given in Eq. (7) or (8). While the determinant of J is zero, the diattenuation is unit and then we define the retardance to zero. It can be shown[6] that a Jones matrix with zero determinant describes a PE whose output light is always at a fixed fully polarized state regardless of the incident polarization state. Such PEs are called polarizers. Thus polarizers are PEs with unit diattenuation.

In above paragraphs we have given unambiguous definition for the diattenuation and retardance in general case. In the following we are going to derive their explicit expressions.

A. Homogeneous and inhomogeneous polarization elements

Again, let J be the Jones matrix of the PE under consideration. Its eigenpolarizations $\hat{E}_q$ and $\hat{E}_r$ satisfy the characteristic equations

$$J\hat{E}_q = \xi_q \hat{E}_q, J\hat{E}_r = \xi_r \hat{E}_r \tag{14}$$

where $\xi_q = |\xi_q| \exp(i\delta_q)$ and $\xi_r = |\xi_r| \exp(i\delta_r)$ are the eigenvalues of the Jones matrix. For an inhomogeneous PE, understandably, the inner product of eigenpolarizations will be an interesting quantity. We define the parameter η as

$$\eta = |\hat{E}_q^* \hat{E}_r|, \quad 1 > \eta \geq 0. \tag{15}$$

For a homogeneous PE, η is zero. Thus η can be regarded as a indicator of inhomogeneity, and, for convenience, we call it inhomogeneity. Its geometrical meaning becomes transparent on the Poincare sphere. Let s_q and s_r be the eigenpolarizations represented the on the Poincare sphere, and χ_{qr} be the half of angle subtended between them (Fig. 1). It can be shown[8] that

$$\eta = \cos \chi_{qr}. \tag{16}$$

Thus $\cos^{-1} \eta$ can be regarded as the "angle" between eigenpolarizations. χ_{qr} is $\frac{\pi}{2}$ for a homogeneous PE and χ_{qr} is less than $\frac{\pi}{2}$ for an inhomogeneous PE. Note that, from Eq. (16), η^2 is equal to the "similarity factor" of the eigenpolarizations introduced in Ref. 9.

Since T_{max} and T_{min} are the extreme values of intensity transmittance, by differentiating Eq. (9) it can be found that they satisfy the following quadratic equation

$$(1-\eta^2)T^2 + [|\xi_q|^2 + |\xi_r|^2 - \eta^2(\xi_q^* \xi_r + \xi_q \xi_r^*)]T + (1-\eta^2)|\xi_q|^2|\xi_r|^2 = 0, \tag{18}$$

i.e.

$$T_{max} T_{min} = |\xi_q|^2 |\xi_r|^2, \tag{19}$$

$$T_{max} + T_{min} = \frac{|\xi_q|^2 + |\xi_r|^2 - \eta^2(\xi_q^* \xi_r + \xi_q \xi_r^*)}{1-\eta^2}. \tag{20}$$

From Eqs. (10), (19) and (20), the expression for the diattenuation are:

$$\mathcal{D} = \left\{ 1 - \frac{4(1-\eta^2)^2 |\xi_q|^2 |\xi_r|^2}{\left[|\xi_q|^2 - |\xi_r|^2 - \eta^2(\xi_q^* \xi_r + \xi_q \xi_r^*)\right]^2} \right\}^{1/2}. \tag{21}$$

The combination of Eqs. (8), (13), (19) and (20) yields the expression for the retardance:

$$\mathcal{R} = 2\cos^{-1}\left\{ \left[\frac{(1-\eta^2)(|\xi_q| + |\xi_r|)^2}{(|\xi_q| - |\xi_r|)^2 - \eta^2(2|\xi_q||\xi_r| + \xi_q^* \xi_r + \xi_q \xi_r^*)} \right]^{1/2} \cos\frac{|\delta_q - \delta_r|}{2} \right\}. \tag{22}$$

While inhomogeneity is zero, Eqs. (21) and (22), as expected, reduce to Eqs. (1) and (2), respectively. From Eq. (21), diattenuation can not be zero except for retarder. From Eq. (22), retardance is zero only for diattenuator and polarizer. In other words, inhomogeneous PEs have both nonzero diattenuation and nonzero retardance, except for inhomogeneous polarizers whose retardance is defined to be zero.

B. *Degenerate polarizing elements*

In case that the Jones matrix $\mathbf{J}$ describes a degenerate PE, we need to redefine η as

$$|(\mathbf{J} - \xi_q \mathbf{I})\hat{\mathbf{F}}_q| = \eta, \quad \eta \neq 0 \tag{23}$$

where $\mathbf{I}$ is the identity matrix, $\xi_q (= \xi_r)$ is the only eigenvalue, and $\hat{\mathbf{F}}_q$ is an arbitrary Jones vector normal to the eigenpolarization $\hat{\mathbf{E}}_q$. It is shown in appendix A that $\mathbf{J}$ can

be transformed into

$$\begin{bmatrix} \xi_q & \eta \\ 0 & \xi_q \end{bmatrix} \tag{24}$$

by an unitary change of basis. From the representation in Eq. (24), η can be regarded as the "coupling" between two orthogonal polarization states $\hat{E}_q$ and $\hat{F}_q$.

With η given above, the T_{max} and T_{min} are given by

$$T_{max} = |\xi_q|^2 + \frac{\eta^2}{2} + \eta\sqrt{|\xi_q|^2 + \frac{\eta^2}{4}} \tag{25}$$

$$T_{min} = |\xi_q|^2 + \frac{\eta^2}{2} - \eta\sqrt{|\xi_q|^2 + \frac{\eta^2}{4}}. \tag{26}$$

And then we have

$$\mathcal{D} = \frac{\eta}{2|\xi_q|^2 - \eta^2}\sqrt{+\,|\xi_q|^2 + \eta^2}, \tag{27}$$

$$\mathcal{R} = 2\cos^{-1}\frac{|\xi_q|}{\sqrt{|\xi_q|^2 + \frac{\eta^2}{4}}}. \tag{28}$$

If η is zero, the Jones matrix is proportional to the identity matrix and, as expected, Eqs. (27) and (28) give zero diattenuation and zero retardance.

C. Classification of polarization elements

It is important to realize that both eigenvalues and η of a PE are unitary invariants, so are diattenuation and retardance. The following theorem indicates that the eigenvalues and η are also the only three unitary invariants a Jones matrix can have.

<u>The Unitary Equivalence Theorem:</u> Two PEs have the same eigenvalues and η if and

only if they are unitarily equivalent, i.e., there exists an unitary matrix $\mathbf{U}$ such that

$$\mathbf{J}_2 = \mathbf{U}^* \mathbf{J}_1 \mathbf{U} \tag{29}$$

where $\mathbf{J}_1$ and $\mathbf{J}_2$ are their Jones matrices.

A proof is given in Appendix B. The transform in Eq. (29) is the same as a base changing transform, and then we can say that two unitarily equivalent Jones matrices, in fact, represent a same PE but in different basis. Thus PEs can be classified into sets of unitarily equivalent PEs with same parameters $(\xi_q, \xi_r; \eta)$. The order between ξ_q and ξ_r are irrelevant. The set of unitarily equivalent PEs with $(\xi_q, \xi_r; \eta)$ will be denoted by PE$(\xi_q, \xi_r; \eta)$. It should be noted that PEs of the set PE$(\xi_q, \xi_r; \eta)$ also have the same diattenuation and retardance. The classification scheme discussed here is simple and elegant.

Let $\Sigma|J|^2 = |j_{11}|^2 + |j_{12}|^2 + |j_{21}|^2 + |j_{22}|^2$ where $j_{mn}(m, n = 1, 2)$ are matrix elements of the Jones matrix of the PE. It is easy to see that $(\Sigma|J|^2, Tr\,\mathbf{J}, Det\,\mathbf{J})$ are another important set of unitary invariants which will be useful in the following discussion.

3. DATA REDUCTION

When the Jones matrix $\mathbf{J}$ of a PE is known, the eigenvalues and η are given by

$$\xi_q, \xi_r = \frac{Tr\,\mathbf{J}}{2} \pm \frac{1}{2}\sqrt{(Tr\,\mathbf{J})^2 - 4\,Det\,\mathbf{J}}, \tag{30}$$

$$\eta^2 = \frac{\Sigma|J|^2 - \frac{1}{2}(|(Tr\,\mathbf{J})^2| + |(Tr\,\mathbf{J})^2 - 4\,Det\,\mathbf{J}|)}{\Sigma|J|^2 - \frac{1}{2}(|(Tr\,\mathbf{J})^2| - |(Tr\,\mathbf{J})^2 - 4\,Det\,\mathbf{J}|)}, \tag{31}$$

or, for a degenerate PE,

$$\eta = |J_{12}| + |J_{21}|. \tag{32}$$

The diattenuation and retardance can be obtained by Eqs. (21) and (22) or, for a degenerate PE, by Eqs. (27) and (28). The following equations provide another way to calculate the diattenuation and retardance:

$$\mathcal{D} = \sqrt{1 - \frac{4|Det\,J|^2}{(\Sigma|J|^2)^2}} \tag{33}$$ [3]

$$\mathcal{R} = 2\cos^{-1}\frac{\left|Tr\,J + \frac{Det\,J}{|Det\,J|}(Tr\,J)^*\right|}{2\sqrt{\Sigma|J|^2 + 2|Det\,J|}}. \tag{34}$$

For each nondepolarizing Mueller matrix, its corresponding Jones matrix can readily be obtained up to an unimportant total phase by the following equations[11]:

$$|J_{11}| = \sqrt{\frac{m_{11} + m_{12} + m_{21} + m_{22}}{2}} \tag{35}$$

$$|J_{12}| = \sqrt{\frac{m_{11} - m_{12} + m_{21} - m_{22}}{2}} \tag{36}$$

$$|J_{21}| = \sqrt{\frac{m_{11} + m_{12} - m_{21} - m_{22}}{2}} \tag{37}$$

$$|J_{22}| = \sqrt{\frac{m_{11} - m_{12} - m_{21} + m_{22}}{2}} \tag{38}$$

$$\phi_{12} - \phi_{11} = \tan^{-1}\left(\frac{-m_{14} - m_{24}}{m_{13} + m_{23}}\right) \tag{39}$$

$$\phi_{21} - \phi_{11} = \tan^{-1}\left(\frac{m_{41} + m_{42}}{m_{31} + m_{32}}\right) \tag{40}$$

$$\phi_{22} - \phi_{11} = \tan^{-1}\left(\frac{m_{43} - m_{34}}{m_{33} + m_{34}}\right) \tag{41}$$

in which ϕ_{mn} (m,n = 1,2) denote the phases of J_{mn}. Thus the data reduction for Mueller matrices is the same as for Jones matrices and Eqs. (30)-(34) are still applicable. From

the definition and property of diattenuation, the following equation also gives diattenuation:

$$\mathcal{D} = \frac{\sqrt{m_{12}^2 + m_{13}^2 + m_{14}^2}}{m_{11}} = \frac{\sqrt{m_{21}^2 + m_{31}^2 + m_{41}^2}}{m_{11}} \tag{42}$$

4. CONCLUSION

The definition of diattenuation and retardance have been generalized to include inhomogeneous PEs, by applying polar decomposition theorem. Their explicit expressions in terms of unitary invariants- eigenvalues and η, have been obtained. These expressions can readily be used to perform data reduction on Jones and nondepolarizing Mueller matrices. The new unitary invariant η we introduced has been proved to be useful. Our results also include the classification of PEs by their unitary invariants. Diattenuation, retardance and inhomogeneity are useful quantities that characterize polarization properties of the PE. To treat depolarizing elements modification and further generalization of the approach in this paper are needed.

APPENDIX A: A Proof of the Property of Degenerate Polarizing Elements

According to the Cayley-Hamilton theorem[7], we have

$$(\mathbf{J} - \xi_q \mathbf{I})^2 = 0. \tag{43}$$

Thus, for arbitrary vector $\mathbf{E}$, $(\mathbf{J} - \xi_q \mathbf{I})\mathbf{E}$ is always parallel to $\hat{\mathbf{E}}_q$. By properly choosing the absolute phase of $\hat{\mathbf{F}}_q$, the following equation will hold:

$$\mathbf{J}\hat{\mathbf{F}}_q = \xi_q \hat{\mathbf{F}}_q + \eta \hat{\mathbf{E}}_q. \tag{44}$$

$\hat{\mathbf{E}}_q$ and $\hat{\mathbf{F}}_q$ are orthonormal Jones vectors, so the matrix $[\hat{\mathbf{E}}_q, \hat{\mathbf{F}}_q]$, with columns equal to $\hat{\mathbf{E}}_q$ and $\hat{\mathbf{F}}_q$, is an unitary matrix. By some vector and matrix manipulation, we find

that

$$
\begin{aligned}
&[\hat{E}_q, \hat{F}_q]^{*} J [\hat{E}_q, \hat{F}_q] \\
&= [\hat{E}_q, \hat{F}]^{*} [\xi_q \hat{E}_q, \xi_q \hat{F} + \eta \hat{E}_q] \\
&= \begin{bmatrix} \xi_q & \eta \\ 0 & \xi_q \end{bmatrix}.
\end{aligned}
\tag{45}
$$

APPENDIX B: A Proof of the Unitary Equivalence Theorem

Since the Jones matrix of any degenerate PE can be transformed into Eq. (24), obviously this theorem holds for degenerate PEs. Thus we only have to consider PEs with two eigenpolarizations. Since the sufficient condition is quite obvious, we give a proof of the necessary condition only.

Necessary condition: Without losing generality we assume $\xi_{q1} = \xi_{q2}$ and $\xi_{r1} = \xi_{r2}$ where ξ's are the eigenvalues of the two PEs. Let $(\hat{E}_{q1}, \hat{E}_{r1})$ and $(\hat{E}_{q2}, \hat{E}_{r2})$ are the eigenpolarizations of the two PEs. Since the two PEs have the same η, by properly choosing the relative phases we can have

$$\hat{E}_{q1}^{*} \hat{E}_{r1} = \hat{E}_{q2}^{*} \hat{E}_{r2}. \tag{46}$$

There exists a matrix U such that

$$\hat{E}_{q2} = U \hat{E}_{q1}, \; \hat{E}_{r2} = U \hat{E}_{r1} \tag{47}$$

From Eq. (46), U is unitary. From Eq. (47), it is the desired unitary transform matrix in Eq. (29).

REFERENCES

1. R. A. Chipman, "Polarization analysis of optical systems," Opt. Eng. **28**, 90-99 (1989).

2. W. A. Shurcliff, *Polarized Light* (Harvard U. Press, Cambridge, Mass. 1962).

3. R. M. A. Azzam and N. M. Bashara, *Ellipsometry and Polarized Light* (North Holland, 1977)

4. J. J. Gil and E. Bernabeu, "Obtainment of the polarizing and retardation parameters of a non-depolarizing optical system from the polar decomposition of its Mueller matrix," Optik, **76**, 67-71 (1987).

5. J. Byrne, "A classification of electron and optical polarization transfer matrices," J. Phys. **B4**, 940-953 (1971).

6. C. Whitney, "Pauli-algebraic operators in polarization optics," J. Opt. Soc. Am. **61**, 1207-1213 (1971).

7. P. Lancaster and M. Tismenetsky, *The Theory of Matrices* , 2nd edition (Prentice-Hall, 1985).

8. G. N. Ramachandran and S. Ramaseshan, "Crystal optics," vol. XXV/1, *Handbuch der Physik*, ed. by S. Flugge (Springer-Verlag, 1961).

9. S. Pancharatnam, "Light Propagation in absorbing crystals possessing optical activity," Proc. India Acad. Sci. **46A**, 28--302 (1957).

10. E. L. O'Neil, *Introduction to Statistical Optics* (Addison-Wesley, 1963).

11. D. Goldstein, Ph.D. dissertation, University of Alabama in Huntsville (1990).

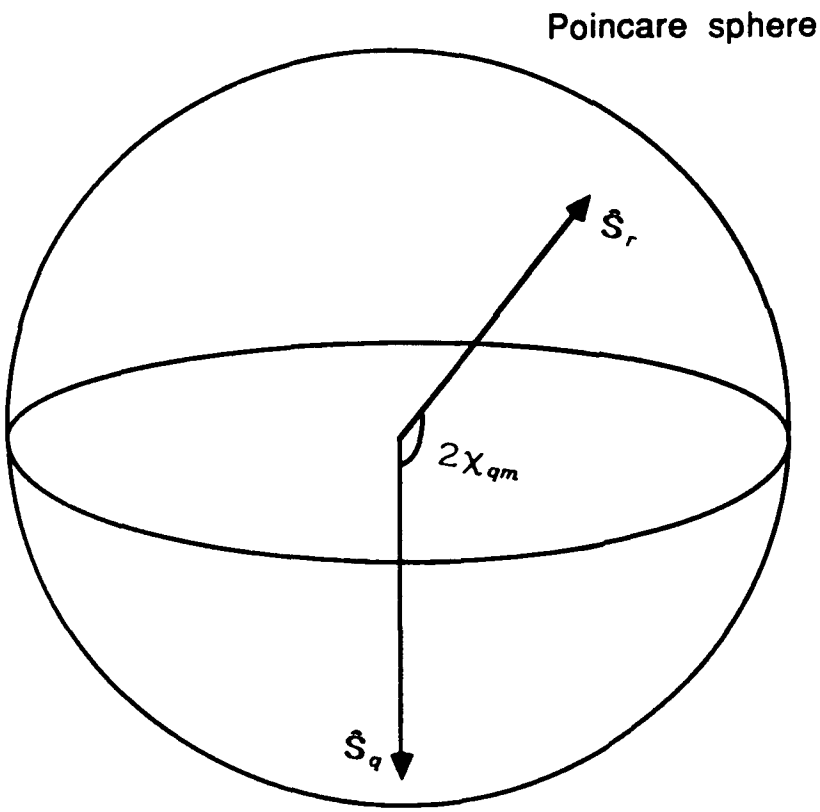


Fig. 1

STUDIES ON JONES CALCULUS

Shih-Yau Lu
Russell A. Chipman
Department of Physics
University of Alabama in Huntsville

§I On Homogeneous Retarders and Diattenuators :

An nxn matrix is called normal, when it has n orthogonal eigenvectors. A homogeneous polarization element has a normal Jones matrix. Specially, a homogeneous retarder (HR) is proportional to a unitary matrix and a homogeneous diattenuator (HD) is proportional to a nonnegative Hermitian matrix. We find some theorems on unitary and Hermitian matrices can readily be applied to HR's and HD's. A few examples of such application are listed below.

(i) A matrix **M** is normal if and only if $\mathbf{M}\mathbf{M}^\dagger = \mathbf{M}^\dagger\mathbf{M}$.

(ii) Any 2x2 Hermitian matrix **H** can be be expanded as
$\mathbf{H} = c_0\sigma_0+c_1\sigma_1+c_2\sigma_2+c_3\sigma_5$
where c's are real.

(iii) For any unitary matrix **U** we can find a Hermitian matrix **H** such that
$\mathbf{U} = \exp(i\mathbf{H})$.

(iv) The product of unitary matrices is unitary.

(v) Any matrix **M** can be expressed in form, the polar representation,
$\mathbf{M} = \mathbf{H}\mathbf{U} = \mathbf{U}'\mathbf{H}'$
and
$\mathbf{H} = (\mathbf{M}\mathbf{M}^\dagger)^{1/2}$ $\mathbf{H}' = (\mathbf{M}^\dagger\mathbf{M})^{1/2}$
$\mathbf{U} = \mathbf{H}^{-1}\mathbf{M}$ $\mathbf{U}' = \mathbf{M}\mathbf{H}'^{-1}$ (if $\det(\mathbf{M})\neq 0$)
where **U**'s and **H**'s are unitary and nonnegative Hermitian matrices respectively.

(i) A polarization element **J** is homogeneous if and only if
$\mathbf{J}\mathbf{J}^\dagger = \mathbf{J}^\dagger\mathbf{J}$.

(ii) The most general form for the HD is
$\mathbf{J}_{HD} = a(c_0\sigma_0+c_1\sigma_1+c_2\sigma_2+c_3\sigma_3)$
where a can be complex and c's are real and
$c_0^2 \geq c_1^2+c_2^2+c_3^2$.

(iii) The most general form for the HR is
$\mathbf{J}_{NR} = a\exp[i(c_0\sigma_0+c_1\sigma_1+c_2\sigma_2+c_3\sigma_3)]$
where a can be complex and c's are real.

(iv) The cascade of HR's is still a HR.

(v) Any Jones matrix **J** can be expressed in the form
$\mathbf{J} = \mathbf{J}_{HD}\mathbf{J}_{HR} = \mathbf{J}'_{HR}\mathbf{J}'_{HD}$
which means any Jones matrix can be realized by only one HR and HD.

Not all retarders and diattenuators are homogeneous. The most general form for the retarder is

$$\mathbf{J}_R \propto \mathbf{A}\begin{bmatrix} e^{i\alpha} & 0 \\ 0 & e^{i\beta} \end{bmatrix}\mathbf{A}^{-1}$$

and for the diattenuator is

$$\mathbf{J}_D \propto \mathbf{A}\begin{bmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{bmatrix}\mathbf{A}^{-1}$$

where **A** is an arbitrary nonsingular matrix, α and β are real, γ_1 and γ_2 are positive real. Note that these forms can not represent the "diattenuators" which are similar to $\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}$.

§II On the Couplance of Inhomogeneous Polarization Elements :

In the following discussion, the polarization element has the Jones matrix

$$\mathbf{J} = c_0\sigma_0+c_1\sigma_1+c_2\sigma_2+c_3\sigma_3$$

and two normalized eigenpolarization states $\mathbf{E}_q$ and $\mathbf{E}_r$.

We learned that, for homogeneous polarization elements, the maximum coupling occurs when the input state **E** is

$$\mathbf{E} = (\mathbf{E}_q + e^{i\delta}\mathbf{E}_r)/\sqrt{2} \qquad (2\text{-}1)$$

where δ is an arbitrary real number, and the couplance C is

$$C = |c_1^2+c_2^2+c_3^2|/(|c_0^2|+|c_1^2+c_2^2+c_3^2|). \qquad (2\text{-}2)$$

The objective of this note is to discuss the maximum coupling input state and the couplance of the inhomogeneous polarization elements.

Suppose the **J** is an inhomogeneous polarization element with two eigenpolarization states. We let

$$\mathbf{J}_\sigma = c_1\sigma_1+c_2\sigma_2+c_3\sigma_3$$

which is the "effective" part of the polarization element, and thus

$$\mathbf{J} = c_0\sigma_0+\mathbf{J}_\sigma .$$

The $\mathbf{J}_\sigma$ has the same eigenpolarization states as the **J**, i.e.,

$\mathbf{J}_\sigma\mathbf{E}_q = \xi\mathbf{E}_q \qquad \mathbf{J}_\sigma\mathbf{E}_r = -\xi\mathbf{E}_r$

where $\xi = (c_1^2+c_2^2+c_3^2)^{1/2}$. Since $\mathbf{JE} = c_o\mathbf{E}+\mathbf{J}_\sigma\mathbf{E}$, the maximum coupling must occur when

$$\mathbf{E}^\dagger(\mathbf{J}_\sigma\mathbf{E}) = 0 \tag{2-3}$$

and $|\mathbf{J}_\sigma\mathbf{E}|$ is at its maxima. Fortunately these two conditions can be satisfied simultaneously. We can express $\mathbf{E}$ as

$$\mathbf{E} = a\mathbf{E}_q+ b\mathbf{E}_r \tag{2-4}$$

and easily get

$$\mathbf{J}_\sigma\mathbf{E} = \xi(a\mathbf{E}_q - b\mathbf{E}_r). \tag{2-5}$$

It follows from eq.(2-3,4,5) that

$$[a\mathbf{E}_q+ b\mathbf{E}_r]^\dagger[\xi(a\mathbf{E}_q - b\mathbf{E}_r)] =0$$

and then

$$\xi\{|a|^2 -|b|^2 -2i[\mathrm{Im}(a^*b\mathbf{E}_q^\dagger\mathbf{E}_r)]\} = 0.$$

Suppose $\xi \neq 0$, otherwise $\mathbf{J} = c_o\sigma_o$ and no coupling can occur. We have

$$|a| = |b| \quad \text{and} \quad \mathrm{Arg}(a^*b) = \pi -\mathrm{Arg}(\mathbf{E}_q^\dagger\mathbf{E}_r) \text{ or } -\mathrm{Arg}(\mathbf{E}_q^\dagger\mathbf{E}_r).$$

Besides, to make $|\mathbf{J}_\sigma\mathbf{E}|$ maximum, we have to chose $\mathrm{Arg}(a^*b) = \pi -\mathrm{Arg}(\mathbf{E}_q^\dagger\mathbf{E}_r)$. Therefore, the maximum coupling input state $\mathbf{E}$ is

$$\mathbf{E} = (\mathbf{E}_q- e^{i\alpha}\mathbf{E}_r)/(2 - 2|\mathbf{E}_q^\dagger\mathbf{E}_r|)^{1/2} \quad \text{and} \quad \alpha = \pi -\mathrm{Arg}(\mathbf{E}_q^\dagger\mathbf{E}_r).$$

The relative phase can be arbitrary only for homogeneous case. When the maximum coupling occurs, the fraction intensity in input state is

$$I_E = (c_o\mathbf{E})^\dagger(c_o\mathbf{E}) = |c_o^2|$$

and the fraction intensity in orthogonal state is

$$I_{E\perp} = (\mathbf{J}_\sigma\mathbf{E})^\dagger(\mathbf{J}_\sigma\mathbf{E}) = \kappa|c_1^2+c_2^2+c_3^2|$$

where $\kappa = (1+|\mathbf{E}_q^\dagger\mathbf{E}_r|)/(1-|\mathbf{E}_q^\dagger\mathbf{E}_r|) \geq 1$. Finally we get the couplance C

$$C = I_{E\perp}/(I_E+I_{E\perp}) = \kappa|c_1^2+c_2^2+c_3^2|/(|c_o^2|+\kappa|c_1^2+c_2^2+c_3^2|).$$

We conclude that, for any polarization element, homogeneous or not, with two eigenpolarization states, the maximum coupling occurs when input state is an equal mixture of the eigenpolarization states.

§III On the "Reversibility" of Jones Matrices

A polarization element or Jones matrix is said to be reversible if it can be realized at least in one way that works for light traveled in both direction. For convenience, we use superscript "~" to denote the Jones matrix in which light travels in the opposite direction. Clearly, if

$$\mathbf{J} = \mathbf{J}_1\mathbf{J}_2\cdots\mathbf{J}_n$$

then

$$\tilde{\mathbf{J}} = \tilde{\mathbf{J}}_n\cdots\tilde{\mathbf{J}}_2\tilde{\mathbf{J}}_1.$$

Let $\mathbf{J}_L(\xi_x,\xi_y)$ and $\mathbf{J}_C(\alpha)$ represent the Jones matrices which has the following forms and properties

$$\mathbf{J}_L(\xi_x,\xi_y) = \begin{bmatrix} \xi_x & 0 \\ 0 & \xi_y \end{bmatrix} \qquad \mathbf{J}_L(\xi_x,\xi_y)\mathbf{E}_x = \xi_x\mathbf{E}_x \qquad \mathbf{J}_L(\xi_x,\xi_y)\mathbf{E}_y = \xi_y\mathbf{E}_y$$

$$\mathbf{J}_C(\alpha) = \begin{bmatrix} \cosh\alpha & i\sinh\alpha \\ -i\sinh\alpha & \cosh\alpha \end{bmatrix} \qquad \mathbf{J}_C(\alpha)\mathbf{E}_R = e^{\alpha}\mathbf{E}_R \qquad \mathbf{J}_C(\alpha)\mathbf{E}_L = e^{-\alpha}\mathbf{E}_L$$

where ξ_x,ξ_y and α are all complex. Besides we define the $\mathbf{J}_C(\infty)$ to be a ideal RHC polarizer and $\mathbf{J}_C(-\infty)$ a ideal LHC polarizer. The $\mathbf{J}_L$ is just a cascade of LD(0°) and LR(0°). Obviously the $\mathbf{J}_L$ behaves similarly in reverse, i.e.,

$$\tilde{\mathbf{J}}_L = \mathbf{J}_L.$$

The $\mathbf{J}_C$ is a cascade of CD and CR, and it can be realized by the optically active material. We do not have to discuss the "reversibility" of optically active material, because $\mathbf{J}_C$ can also be realized by

$$\begin{aligned}
\mathbf{J}_C &= \mathrm{QWLR}(90^\circ)\ [\mathbf{J}_L(e^{\alpha},e^{-\alpha}) \text{ at } 45^\circ]\ \mathrm{QWLR}(0^\circ) \\
&= \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix} \begin{bmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix} \\
&= \begin{bmatrix} \cosh\alpha & i\sinh\alpha \\ -i\sinh\alpha & \cosh\alpha \end{bmatrix}.
\end{aligned}$$

By this realization $\mathbf{J}_C$ is reversible, i.e.,

$$\begin{aligned}
\tilde{\mathbf{J}}_C &= \mathrm{QWLR}(0^\circ)\ [\mathbf{J}_L(e^{\alpha},e^{-\alpha}) \text{ at } -45^\circ]\ \mathrm{QWLR}(90^\circ) \\
&= \begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix} \begin{bmatrix} \cosh\alpha & -\sinh\alpha \\ -\sinh\alpha & \cosh\alpha \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix} \\
&= \begin{bmatrix} \cosh\alpha & i\sinh\alpha \\ -i\sinh\alpha & \cosh\alpha \end{bmatrix} \\
&= \mathbf{J}_C.
\end{aligned}$$

We think that the $\mathbf{J}_L$ and $\mathbf{J}_C$ are the most elementary and important reversible Jones matrices. They are very useful in the discussion of the "reversibility" of Jones matrices.

Any Jones matrix $\mathbf{J}$ can be expressed in the form

$$\mathbf{J} = \mathbf{J}_C(\alpha)\mathbf{J}_L(\lambda_x,\lambda_y)\mathbf{J}_C(\beta).$$

Since we can solve the following four equations

$$J_{11}+J22 = (\lambda_x+\lambda_y)\cosh(\alpha+\beta) \qquad (3\text{-}1a)$$

$$J_{11}-J22 = (\lambda_x-\lambda_y)\cosh(\alpha-\beta) \qquad (3\text{-}1b)$$

$$J_{12}+J21 = -i\ (\lambda_x-\lambda_y)\sinh(\alpha-\beta) \qquad (3\text{-}1c)$$

$$J_{12}-J21 = i\ (\lambda_x+\lambda_y)\sinh(\alpha+\beta) \qquad (3\text{-}1d)$$

to get the four unknowns, α, β, λ_x and λ_y, at leat one set of solutions exist. From eq.(3-1c), the Jones matrix $\mathbf{J}$ with $J_{12}+J_{21} = 0$ has the symmetrical expression

$$\mathbf{J} = \mathbf{J}_C(\alpha)\mathbf{J}_L(\lambda_x,\lambda_y)\mathbf{J}_C(\alpha)$$

and thus

$$\mathbf{J} = \tilde{\mathbf{J}}_C(\alpha)\tilde{\mathbf{J}}_L(\lambda_x,\lambda_y)\tilde{\mathbf{J}}_C(\alpha) = \mathbf{J}_C(\alpha)\mathbf{J}_L(\lambda_x,\lambda_y)\mathbf{J}_C(\alpha) = \mathbf{J}$$

It is found that $J_{12}+J_{21} = 0$ is the sufficient condition for the $\mathbf{J}$ to be reversible.

We can go further. Clearly the most general forms of the reversible Jones matrices are

$$\cdots\cdots\mathbf{J}_L''\mathbf{J}_C'\mathbf{J}_L\mathbf{J}_C'\mathbf{J}_L''\cdots\cdots$$

and

$$\cdots\cdots\mathbf{J}_C''\mathbf{J}_L'\mathbf{J}_C\mathbf{J}_L'\mathbf{J}_C''\cdots\cdots.$$

As $(J_L)_{12}+(J_L)_{21} = 0$ and $(J_C)_{12}+(J_C)_{21} = 0$, it follows that

$$(\mathbf{J}_C'\mathbf{J}_L\mathbf{J}_C')_{12}+(\mathbf{J}_C'\mathbf{J}_L\mathbf{J}_C')_{21} = 0$$

and

$$(\mathbf{J}_L'\mathbf{J}_C\mathbf{J}_L')_{12}+(\mathbf{J}_L'\mathbf{J}_C\mathbf{J}_L')_{21} = 0.$$

Just by repeating the same argument it is not difficulty to get that

$$(\cdots\cdots\mathbf{J}_L''\mathbf{J}_C'\mathbf{J}_L\mathbf{J}_C'\mathbf{J}_L''\cdots\cdots)_{12}+(\cdots\cdots\mathbf{J}_L''\mathbf{J}_C'\mathbf{J}_L\mathbf{J}_C'\mathbf{J}_L''\cdots\cdots)_{21} = 0$$

and

$$(\cdots\cdots\mathbf{J}_C''\mathbf{J}_L'\mathbf{J}_C\mathbf{J}_L'\mathbf{J}_C''\cdots\cdots)_{12}+(\cdots\cdots\mathbf{J}_C''\mathbf{J}_L'\mathbf{J}_C\mathbf{J}_L'\mathbf{J}_C''\cdots\cdots)_{21} = 0.$$

Therefore $J_{12}+J_{21} = 0$ is also the necessary condition for the $\mathbf{J}$ to be reversible.

To get more physical insight we have to go back to the Pauli matrices. Without doubt the σ_0 and σ_1 are reversible. Since the σ_2 is a HWLR(45°), it is not difficulty to understand that it is irreversible and $\tilde{\sigma}_2 = -\sigma_2$. The σ_3 is a HWCR and

$\sigma_3 = J_C(\, i\pi/2)$. Since the J_C is reversible, so is the σ_3. The properties of the Pauli matrices are summarized below.

$$\tilde{\sigma}_o = \sigma_o \quad \tilde{\sigma}_1 = \sigma_1 \quad \tilde{\sigma}_2 = -\sigma_2 \quad \tilde{\sigma}_3 = \sigma_3.$$

Suppose the Jones matrix **J** has the following expression

$$\mathbf{J} = c_o\sigma_o + c_1\sigma_1 + c_2\sigma_2 + c_3\sigma_3,$$

then we have

$$\tilde{\mathbf{J}} = c_o\tilde{\sigma}_o + c_1\tilde{\sigma}_1 + c_2\tilde{\sigma}_2 + c_3\tilde{\sigma}_3 = c_o\sigma_o + c_1\sigma_1 - c_2\sigma_2 + c_3\sigma_3.$$

The similar conclusion follows that **J** is reversible if and only if **J** does not contain σ_2. As the σ_2 is the only irreversible one of the Pauli matrices, the result is quite plausible.

§IV On the Decomposition: $\mathbf{M} = \frac{1}{2}(\mathbf{H''U''}+\mathbf{U''H''})$

This note mainly concerns the decomposition

$$\mathbf{M} = \frac{1}{2}(\mathbf{H''U''}+\mathbf{U''H''}), \qquad (4\text{-}1)$$

for an arbitrary complex matrix **M**, where **H"** and **U"** are nonnegative define Hermitian matrix and unitary matrix respectively. I would like to call the decomposition of Eq. (1) the "symmetric polar decomposition (SPD)" in contrast to the polar decomposition of Eq. (2). It is found that not all matrices have the SPD. Besides **H"** and **U"** have the possibility to be unique only for nonsingular matrices. For convenience, we will use the capital **H** to denote a nonnegative define Hermitian matrix, and the capital **U** for an unitary matrix, and $\mathrm{diag}[d_1,\cdots,d_n]$ for a diagonal matrix with diagonal elements $d_1,\cdots,d_n$.

The Polar Decomposition Theorem will paly a crucial role in the following discussion, so we restate it below. In addition, several useful Propositions are also cited without proof.

<u>The Polar Decomposition Theorem</u>. ([1], [2]) For any matrix **M** there exist unique **H**, **H'** and (not necessarily unique) an **U** such that

$$\mathbf{M} = \mathbf{HU} = \mathbf{UH'} \qquad (4\text{-}2)$$

where **H** and **H'** are given by

$$\mathbf{H} = \sqrt{\mathbf{MM}^\dagger} \qquad \mathbf{H'} = \sqrt{\mathbf{M}^\dagger\mathbf{M}}. \qquad (4\text{-}3)$$

and **U** can be determined by solving the linear equation of Eq. (2) with **H** and **H'** given by Eq. (3). Furthermore when **M** is nonsingular **U** is uniquely given by

$$\mathbf{U} = \mathbf{H}^{-1}\mathbf{M} = \sqrt{\mathbf{MM}^\dagger}^{\,-1}\,\mathbf{M}. \qquad (4\text{-}4)$$

<u>Proposition 1</u>. ([1]) A matrix **M** is normal if and only if **H** and **U** in Eq. (2) commute.

<u>Corollary 1</u>. If **M** is normal, then it can be written as

$$\mathbf{M} = \mathbf{HU} = \mathbf{UH} = \frac{1}{2}(\mathbf{HU}+\mathbf{UH})$$

where **H** and **U** are given by the Polar Representation Theorem.

<u>Proposition 2</u>. If $\mathbf{H}_1$ is a nonnegative define Hermitian matrix, then so are $\mathbf{U}_1\mathbf{H}_1\mathbf{U}_1^{-1}$ and $\omega(\mathbf{H}_1+\mathbf{U}_1\mathbf{H}_1\mathbf{U}_1^{-1})$ where $\mathbf{U}_1$ is an unitary matrix and $\omega \geq 0$.

Proposition 3. Suppose $\mathbf{U}_2$ is an unitary matrix and $\mathbf{H}_2$ is a nonnegative define Hermitian matrix. If the equation

$$\mathbf{X}+\mathbf{U}_2\mathbf{X}\mathbf{U}_2^{-1} = \mathbf{H}_2$$

has a solution $\mathbf{X}$, then there exists a solution $\mathbf{X}$ which is Hermitian.

Proposition 4.[1] The matrix equation

$$\mathbf{AX}+\mathbf{XB} = \mathbf{C}$$

has a unique solution if and only if the matrices $\mathbf{A}$ and $-\mathbf{B}$ have no eigenvalue in common.

If $\mathbf{M}$ is normal, the SPD problem is already answered by the Corollary 1.
(A) $\mathbf{M}$ is normal and nonsingular. By the property of normal matrix, $\mathbf{M}$ can be diagonalized by an unitary matrix $\mathbf{U}_3$, i.e.

$$\mathbf{M} = \mathbf{U}_3 \operatorname{diag}[\rho_1 e^{i\phi_1},\cdots,\rho_n e^{i\phi_n}]\, \mathbf{U}_3^{-1}. \qquad (4\text{-}5)$$

Let

$$\mathbf{H}'' = \mathbf{U}_3 \operatorname{diag}[\rho_1,\cdots,\rho_n]\, \mathbf{U}_3^{-1} \qquad (4\text{-}6)$$

and

$$\mathbf{U}'' = \mathbf{U}_3 \operatorname{diag}[e^{i\phi_1},\cdots,e^{i\phi_n}]\, \mathbf{U}_3^{-1}. \qquad (4\text{-}7)$$

Thus $\mathbf{M}$ has the following expression

$$\mathbf{M} = \mathbf{H}''\mathbf{U}'' = \mathbf{U}''\mathbf{H}'' = \frac{1}{2}(\mathbf{H}''\mathbf{U}''+\mathbf{U}''\mathbf{H}'')$$

which is the desired SPD.
(B) $\mathbf{M}$ is normal but singular. $\mathbf{M}$ can be written as

$$\mathbf{M} = \mathbf{U}_4 \operatorname{diag}[\rho_1 e^{i\phi_1},\cdots,\rho_m e^{i\phi_m},0,\cdots,0]\, \mathbf{U}_4^{-1}.$$

Similarily let

$$\mathbf{H}'' = \mathbf{U}_4 \operatorname{diag}[\rho_1,\cdots,\rho_m,0,\cdots,0]\, \mathbf{U}_4^{-1} \qquad (4\text{-}8)$$

and

$$\mathbf{U}'' = \mathbf{U}_4 \operatorname{diag}[e^{i\phi_1},\cdots,e^{i\phi_m}, e^{i\lambda_{m+1}},\cdots, e^{i\lambda_n}]\, \mathbf{U}_4^{-1} \qquad (4\text{-}9)$$

where $0 \le \lambda_{m+1},\cdots,\lambda_n < 2\pi$. It follows that

$$\mathbf{M} = \mathbf{H}''\mathbf{U}'' = \mathbf{U}''\mathbf{H}'' = \frac{1}{2}(\mathbf{H}''\mathbf{U}''+\mathbf{U}''\mathbf{H}'').$$

Note that, because $\mathbf{M}$ is singular, $\mathbf{U}''$ is not unique.

Here we consider the most general case. Suppose the SPD of Eq. (1) holds for arbitrary $\mathbf{M}$. First we rewrite Eq. (1) as

$$\mathbf{M} = \{\frac{1}{2}(\mathbf{H}''+\mathbf{U}''\mathbf{H}''\mathbf{U}''^{-1})\}\mathbf{U}''. \qquad (4\text{-}10)$$

Since we suppose **H"** is a nonnegative a nonnegative, $\frac{1}{2}(\mathbf{H''}+\mathbf{U''H''U''}^{-1})$ is also a nonnegative Hermitian matrix by the Proposition 2. Applying the Polar Decomposition Theorem, it follows that **U"** can be determined by the linear equation

$$\mathbf{M} = \sqrt{\mathbf{MM}^{\dagger}}\ \mathbf{U''} \qquad (4\text{-}11)$$

or by Eq. (4-4) when **M** is nonsingular. Note that the **U"** is the same as **U** of the polar decomposition. After we get **U"**, **H"** can be found by the linear equation

$$\frac{1}{2}(\mathbf{H''}+\mathbf{U''H''U''}^{-1}) = \sqrt{\mathbf{MM}^{\dagger}} \qquad (4\text{-}12)$$

or, equally, by

$$\mathbf{H''U''}+\mathbf{U''H''} = 2\,\mathbf{M}. \qquad (4\text{-}13)$$

Since the Eq. (4-12) or (4-13) might ot have a nonnegatively Hermitian solution **H"**, the SPD is not always possible. In order to know wether there is a nonegatively Hermitian solution or not, we have to,in general, exame Eq. (4-12) or (4-13) case by case. To my konwledge there is no principle to follow.

The above-mentioned results can be readily applied to Jones matrices by the following correspondence

M	----	**J**
H"	----	$\mathbf{J}_{HD}$
U"	----	$\mathbf{J}_{HR}$

Since Jones matrices have very low dimension (in fact, 2-D), the SPD problem is much simpler and $\mathbf{J}_{HD}$ can be written in a closed form

$$\mathbf{J}_{HD} = (\mathbf{J}+\det(\mathbf{J}_{HR})\mathbf{J}^{\dagger})/\mathrm{Tr}(\mathbf{J}_{HR}). \qquad (4\text{-}14)$$

A simple computer program has been written to compute the SPD for real **J**. And it is found that the SPD is valid only for a resticted class of matrices.

<u>Example 1</u>.

$$\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix} = \begin{bmatrix}\sqrt{2} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-e^{i\alpha}}{\sqrt{2}} & \frac{e^{i\alpha}}{\sqrt{2}}\end{bmatrix} = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-e^{i\alpha}}{\sqrt{2}} & \frac{e^{i\alpha}}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix} \qquad 0 \le \alpha < 2\pi$$

$$= \frac{1}{2}\left\{\begin{bmatrix}\sqrt{2} & \frac{\sqrt{2}}{1+e^{i\alpha}}\\ \frac{\sqrt{2}}{1+e^{-i\alpha}} & 0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-e^{i\alpha}}{\sqrt{2}} & \frac{e^{i\alpha}}{\sqrt{2}}\end{bmatrix} + \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{-e^{i\alpha}}{\sqrt{2}} & \frac{e^{i\alpha}}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\sqrt{2} & \frac{\sqrt{2}}{1+e^{i\alpha}}\\ \frac{\sqrt{2}}{1+e^{-i\alpha}} & 0\end{bmatrix}\right\} \alpha \ne \pi.$$

Since $\begin{bmatrix} \sqrt{2} & \frac{\sqrt{2}}{1+e^{i\alpha}} \\ \frac{\sqrt{2}}{1+e^{-i\alpha}} & 0 \end{bmatrix}$ is not a nonnegative define matrix, $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ has no SPD.

REFERENCES

[1] P. Lancaster & M. Tismenetsky, *The theory of Matrices*, 2nd ed. 1985, chap. 5 &12.

[2] J. G. Broida & S. G. Williamson, *A Comprehensive Introduction to Linear Algebra*, Addison-Wesley, 1989, pp.540.

§V On Several Order Independent Equalities

It is well understood that, for any matrices **X** and **Y**, if [**X**,**Y**]=0, then

$$\exp\mathbf{X}\exp\mathbf{Y} = \exp(\mathbf{X}+\mathbf{Y}).$$

This is in general not true when [**X**,**Y**]≠0, and two questions arise naturally. One is whether one can obtain a explicit formula for exp**X**exp**Y**. In other words what is the exponent of the RH side of the equation

$$\exp\mathbf{X}\exp\mathbf{Y} = \exp(?).$$

The second one is what kind of operation makes the following equation hold.

$$\exp\mathbf{X}(?)\exp\mathbf{Y} = \exp(\mathbf{X}+\mathbf{Y})$$

The first question is solved by the following Theorem.

<u>Theorem 1</u>. ([1]) Let **X** and **Y** be two complex matrices. Then we have

$$\exp\mathbf{X}\exp\mathbf{Y} = \exp\{\mathbf{X}+\mathbf{Y}+\frac{1}{2}[\mathbf{X},\mathbf{Y}]+\frac{1}{12}[[\mathbf{X},\mathbf{Y}],\mathbf{Y}]-\frac{1}{12}[\mathbf{X},\mathbf{Y}],\mathbf{X}]+\cdots\cdots\} \quad (5\text{-}1)$$

The high order terms are quite complicated and a recursion formula can be found in [1]. In the theory of Lie groups, Eq.(5-1) is called the Baker-Campell-Hausdroff (BCH) formula.

Corollary 1. ([2],[3])
(i) $\exp(t\mathbf{X})\exp(t\mathbf{Y}) = \exp\{ t(\mathbf{X}+\mathbf{Y}) + \mathbf{O}(t^2)\}$
(ii) $\exp\{t(\mathbf{X}+\mathbf{Y})\} = \exp(t\mathbf{X})\exp(t\mathbf{Y})\exp(\mathbf{O}(t^2))$
where t is a small real parameter and $\mathbf{O}(t^2)$ denotes a term of order t^2.

Corollary 2.
(i) $\exp(t\mathbf{X}+\mathbf{O}'(t^2)) + \exp(t\mathbf{X}+\mathbf{O}''(t^2)) = \exp(t\mathbf{X}+\mathbf{O}(t^2))$
where t is a small real parameter and, $\mathbf{O}'(t^2)$, $\mathbf{O}''(t^2)$ and $\mathbf{O}(t^2)$ denote terms of order t^2.

The next theorem is a consequence of Corollary 1, and it offers a answer to the second question.

Theorem 2. ([2],[3]) For any two complex matrices **X** and **Y**, we have

$$\{\exp(\mathbf{X}/n)\exp(\mathbf{Y}/n)\}^n \underset{n\to\infty}{=\!=\!=} \exp(\mathbf{X}+\mathbf{Y}). \quad (5\text{-}2)$$

Corollary 3.

(i) $\{\exp(\mathbf{Y}/n)\exp(\mathbf{X}/n)\}^n \underset{n\to\infty}{=\!=\!=} \{\exp(\mathbf{X}/n)\exp(\mathbf{Y}/n)\}^n \underset{n\to\infty}{=\!=\!=} \exp(\mathbf{X}+\mathbf{Y})$

(ii) $\{\exp(\mathbf{X}/n)\otimes\exp(\mathbf{Y}/n)\}^n \underset{n\to\infty}{=\!=\!=} \exp(\mathbf{X}+\mathbf{Y})$ (5-3)

(iii) $\{\exp(\mathbf{X}/n)\exp(\mathbf{Y}/n)\exp(\mathbf{Z}/n)\}^n \underset{n\to\infty}{=\!=\!=} \exp(\mathbf{X}+\mathbf{Y}+\mathbf{Z})$ (5-4)

(iv) $\{\exp(\mathbf{X}/n)\otimes\exp(\mathbf{Y}/n)\otimes\exp(\mathbf{Z}/n)\}^n \underset{n\to\infty}{=\!=\!=} \exp(\mathbf{X}+\mathbf{Y}+\mathbf{Z})$ (5-5)

We let $\exp(\mathbf{X})=\mathbf{A}$ and $\exp(\mathbf{Y})=\mathbf{B}$ and so on, then $\exp(\mathbf{X}/n)=\mathbf{A}^{1/n}$ and $\exp(\mathbf{Y}/n)=\mathbf{B}^{1/n}$, etc. So Theorem 1 and Corollary 4 can be rewritten as

$$\begin{aligned}(\mathbf{A}^{1/n}\mathbf{B}^{1/n})^n &\underset{n\to\infty}{=\!=\!=} (\mathbf{B}^{1/n}\mathbf{A}^{1/n})^n \\ &\underset{n\to\infty}{=\!=\!=} (\mathbf{A}^{1/n}\otimes\mathbf{B}^{1/n})^n \\ &\underset{n\to\infty}{=\!=\!=} \exp(\ln\mathbf{A} + \ln\mathbf{B}) \\ &\underset{n\to\infty}{\not=\!=\!=} \mathbf{A}\otimes\mathbf{B}\end{aligned}$$

and

$$(\mathbf{A}^{1/n}\mathbf{B}^{1/n}\mathbf{C}^{1/n})^n \underset{n\to\infty}{=\!=\!=} (\mathbf{A}^{1/n}\otimes\mathbf{B}^{1/n}\otimes\mathbf{C}^{1/n})^n$$

$$\underset{n\to\infty}{=\!=\!=} \exp(\ln\mathbf{A} + \ln\mathbf{B} + \ln\mathbf{C})$$

$$\underset{n\to\infty}{\neq} (\mathbf{A}\otimes\mathbf{B}\otimes\mathbf{C})$$

The inequality $\exp(\ln\mathbf{A} + \ln\mathbf{B})\neq(\mathbf{A}\otimes\mathbf{B})$ is quite obviously, since

$$(\sigma_1\otimes\sigma_2) = 0 \neq \exp(\ln\sigma_1 + \ln\sigma_2).$$

Therefore, in my opinion, equalities in p.73 & 74 of the lecture note are in general not true.

Since any nonsingular Jones matrix **J** has the exponential representation

$$\mathbf{J} = \exp(d_o\sigma_o+d_1\sigma_1+d_2\sigma_2+d_3\sigma_3), \qquad (5\text{-}6)$$

it can be expressed as

$$\{\exp\{i(d_{o,i}\sigma_o+d_{1,i}\sigma_1+d_{2,i}\sigma_2+d_{3,i}\sigma_3)/n\}\{\exp\{(d_{o,r}\sigma_o+d_{1,r}\sigma_1+d_{2,r}\sigma_2+d_{3,r}\sigma_3)/n\}\}^n \underset{n\to\infty}{=\!=\!=} \mathbf{J} \qquad (5\text{-}7)$$

where $d_{,i}$ and $d_{,r}$ denote the imaginary part and real part of d. It is not difficulty to understand that $\exp i(d_{o,i}\sigma_o+d_{1,i}\sigma_1+d_{2,i}\sigma_2+d_{3,i}\sigma_3)/n$ is a homogeneous retarder and $\exp(d_{o,r}\sigma_o+d_{1,r}\sigma_1+d_{2,r}\sigma_2+d_{3,r}\sigma_3)/n$ is a homogeneous diattenuator. So Eq.(5-7) has the interpretation that any nonsingular **J** is an order independent cascade of infinite number of weak $\mathbf{J}_{HR}$ and $\mathbf{J}_{HD}$.

J can also be expressed as

$$\{\exp(d_o\sigma_o/n)\exp(d_1\sigma_1/n)\exp(d_2\sigma_2/n)\exp(d_3\sigma_3/n)\}^n \underset{n\to\infty}{=\!=\!=} \mathbf{J}. \qquad (5\text{-}8)$$

Eq.(5-8) has a similarly explanation as Eq.(5-7).

Since the exponential form of **J** has lots of order independent properties, I think Eq.(6) is the "cannonical" form for a Jones matrix. Specially,

	cannonical form	
Total Phase and Amplitude	$\exp(d_o\sigma_o)$	
Linear Polarization Element at 0°	$\exp(d_1\sigma_1)$	$\Delta\phi=2d_{1,i}$ $D=\tanh 2d_{1,r}$
Linear Polarization Element at 45°	$\exp(d_2\sigma_2)$	$\Delta\phi=2d_{2,i}$ $D=\tanh 2d_{2,r}$
Circular Polarization Element	$\exp(d_3\sigma_3)$	$\Delta\phi=2d_{3,i}$ $D=\tanh 2d_{3,r}$

The results of Theorem 2 and Corollary 4, I believe, are very useful in modeling polarization elements and studying the intrinsic properties of **J**

REFERENCE

[1] V. S. Varadajan, "Lie Groups, Lie Algebras, and their Representations", Prencite-Hall, 1974, pp. 114-121.

[2] P. M. Cohn, "Lie Groups", CUP, 1957, pp.111-112.

[3] M. Sugiura, "Unitary Representations and Harmonic Analysis", Halsted Press (a division of Jone Wiley & Sons), 1975, pp. 55-60.

§VI On Several Properties of the Canonnical Representations of Jones Matrices

In the following we consider an arbitrary nonsingular Jones matrix **J** of the form

$$\mathbf{J} = c_0\sigma_0+c_1\sigma_1+c_2\sigma_2+c_3\sigma_3 = \exp(d_0\sigma_0+d_1\sigma_1+d_2\sigma_2+d_3\sigma_3)$$

or, in vector notation,

$$\mathbf{J} = c_0\sigma_0+\mathbf{c}^T\sigma = \exp(d_0\sigma_0+\mathbf{d}^T\boldsymbol{\sigma})$$

where $\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$, $\mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$, $\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix}$.

Eigenvalues, eigenvectors, determinant and trace

$$\xi_{q,r} = c_o + \sqrt{\mathbf{c}^T\mathbf{c}} = \exp(d_o + \sqrt{\mathbf{d}^T\mathbf{d}})$$
$$\mathrm{Det}(\mathbf{J}) = c_o^2 - \mathbf{c}^T\mathbf{c} = \exp(2d_o)$$
$$\mathrm{Tr}(\mathbf{J}) = 2c_o = 2\exp(\ d_o)\cosh\sqrt{\mathbf{d}^T\mathbf{d}}\,.$$

The $d_o\sigma_o+\mathbf{d}^T\sigma$ and $c_o\sigma_o+\mathbf{c}^T\sigma$ have the same eigenvector(s).

Rotated polarizer

$$\mathbf{J}(\theta) = \mathbf{R}(\theta)\mathbf{J}\mathbf{R}(-\theta)$$
$$= c_o\sigma_o+(c_1\cos2\theta-c_2\sin2\theta)\sigma_1+(c_1\sin2\theta+c_2\cos2\theta)\sigma_2+c_3\sigma_3$$
$$= \exp\{d_o\sigma_o+(d_1\cos2\theta-d_2\sin2\theta)\sigma_1+(d_1\sin2\theta+d_2\cos2\theta)\sigma_2+d_3\sigma_3\}.$$

Note that c- and d-coefficients are transformed in the same way under rotation.

The relation between c- and d-cofficients

Provided $\mathbf{c}^T\mathbf{c}\neq0$ or $\mathbf{d}^T\mathbf{d}\neq0$, then

$$c_o = \exp(d_o)\cosh\sqrt{\mathbf{d}^T\mathbf{d}}\,, \qquad \mathbf{c} = \exp(d_o)\frac{\sinh\sqrt{\mathbf{d}^T\mathbf{d}}}{\sqrt{\mathbf{d}^T\mathbf{d}}}\mathbf{d}$$

and

$$d_o = \frac{1}{2}\ln(c_o^2 - \mathbf{c}^T\mathbf{c})\,, \qquad \mathbf{d} = \frac{\sinh^{-1}\sqrt{\dfrac{\mathbf{c}^T\mathbf{c}}{c_o^2-\mathbf{c}^T\mathbf{c}}}}{\sqrt{\mathbf{c}^T\mathbf{c}}}\mathbf{c}\,.$$

Note that $\mathbf{c}$ and $\mathbf{d}$ are kind of proportional to each other. If $\mathbf{c}^T\mathbf{c}=0$ or $\mathbf{d}^T\mathbf{d}=0$, then

$$c_o = \exp(d_o)\,, \quad \mathbf{c} = \exp(d_o)\,\mathbf{d}$$

and

$$d_o = \ln(c_o), \qquad \mathbf{d} = \frac{1}{c_o}\,\mathbf{c}\,.$$

If $d_o\sigma_o+\mathbf{d}^T\boldsymbol{\sigma}$ is normal, then $\mathbf{J}$ is homogeneous. If $d_o\sigma_o+\mathbf{d}^T\boldsymbol{\sigma}$ is Hermitian, then $\mathbf{J}$ is a HD. The reverse of the above statement statements are also true provided $\mathbf{J}$ is nonsingular. If $d_o\sigma_o+\mathbf{d}^T\boldsymbol{\sigma}$ is anti-Hermitian if and only if $\mathbf{J}$ is a HR.

For singular $\mathbf{J}$, some of d-coefficients are infinite. I would like to call them the "singular polarization elements", since they have infinitive absorption at some polarization states. The question is how to incorporate the "singular polarization elements" into this canonnical representation scheme.

VII On the states of extreme transmittances

1) An Example

We consider an inhomogeneous polarization elements with Jones matrix

$$\begin{bmatrix} 1 & 1 \\ 0 & 1+i \end{bmatrix}.$$

Its eigenpolarizations are

$$\hat{E}_q = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \hat{H}, \quad \hat{E}_r = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ i \end{bmatrix} = \hat{R}$$

and the corresponding eigenvalues are

$$\xi_q = 1, \quad \xi_r = 1 + i.$$

In addition we have

$$\eta^2 = \frac{1}{2}.$$

The T_{max} and T_{min} are the eigenvalues of $J^* J$. The $\hat{E}_{max}$ and $\hat{E}_{min}$ are the eigenvectors of $J^* J$. Since we have

$$J^* J = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix},$$

it can be found that

$$T_{max,min} = 2 \pm \sqrt{2}$$

and

$$\hat{E}_{max} = (4+2\sqrt{2})^{-1/2}\begin{bmatrix} 1 \\ 1+\sqrt{2} \end{bmatrix}, \quad \hat{E}_{min} = (4-2\sqrt{2})^{-1/2}\begin{bmatrix} 1 \\ 1-\sqrt{2} \end{bmatrix}.$$

Clearly $\hat{E}_q$, $\hat{E}_r$, $\hat{E}_{max}$ and $\hat{E}_{min}$ are not on a great circle.

2) Algebraic Picture

The states of extreme transmittances of PE are important. First they have interesting properties:

$$\hat{E}^*_{max}\hat{E}_{min} = 0 \text{ and } (J\hat{E}_{max})^*(J\hat{E}_{min}) = 0. \tag{1}$$

In addition they play a crucial role in discussing the diattenuation. For homogeneous PE the eigenpolarizations are the states of extreme transmittances. It can be shown that, for inhomogeneous PE,

$$\hat{E}_{max} = a\hat{E}_q + b\hat{E}_r \tag{2}$$

where

$$a/b = -\hat{E}^*_q\hat{E}_r(\xi^*_q\xi_r - T_{max})/(|\xi_q|^2 - T_{max}). \tag{3}$$

A similar formula can be found for $\hat{E}_{min}$. Since $\hat{E}_{min}$ is normal to $\hat{E}_{max}$, we only have to discuss $\hat{E}_{max}$.

Based on the result in Ref. 1, the following theorem can be formulated.

<u>Theorem</u>: Let $\hat{E}_3 = a\hat{E}_1 + b\hat{E}_2$. $\hat{S}_1$, $\hat{S}_2$ and $\hat{S}_3$ are on a great circle in the Poincare sphere if and only if $a^*b\hat{E}^*_1\hat{E}_2$ is real.

Therefore, for inhomogeneous PE, eigenpolarizations and the states of extreme transmittances are on same great circle in the Poincare sphere if and only if $\xi^*_q\xi_r$ is real. We reconsider the example discussed before:

$$J = \begin{bmatrix} 1 & 1 \\ 0 & 1+i \end{bmatrix}. \tag{4}$$

Since we have $\xi_q = 1$ and $\xi_r = 1 + i$, $\xi_q^* \xi_r$ is obviously not real. And then we get the same conclusion that its eigenpolarizations and states of extreme transmittances are not on a same great circle in the Poincare sphere.

3) Geometrical Picture

The geometrical picture is more complicated. $\hat{E}_{max}$ is given by Eqs. (2) and (3), so we obtain

$$\cos\frac{\theta_{qm}}{2} = |\hat{E}_q^* \hat{E}_{max}|$$
$$= \eta \left| |\xi_q|^2 - \xi_q^* \xi_r \right| / \sqrt{\eta^2 |T_{max} - \xi_q^* \xi_r|^2 + (T_{max} - |\xi_q|^2)^2} \tag{5}$$

and

$$\cos\frac{\theta_{rm}}{2} = |\hat{E}_r^* \hat{E}_{max}|$$

$$= \eta \left| (T_{max} - \xi_q^* \xi_r) - (T_{max} - |\xi_q|^2) \right| / \sqrt{\eta^2 |T_{max} - \xi_q^* \xi_r|^2 + (T_{max} - |\xi_q|^2)^2} \tag{6}$$

where

$$\eta = |\hat{E}_q^* \hat{E}_r| . \tag{7}$$

θ_{qm} is the angle between $\hat{s}_q$ and $\hat{s}_{max}$ in the Poincare sphere. θ_{rm} is defined similarly. The stokes vectors which have a fixed angle with $\hat{s}_q$ (or $\hat{s}_r$) constitute a circle in the Poincare sphere. From Eqs. (5) and (6) we see that $\hat{s}_{max}$ is located at the intersection of two circles (Fig. 1). Note that two circles can have two intersection points. If $\mathrm{Im}(\xi_q^* \xi_r) > 0$, then $\hat{s}_{max}$ is chosen such that the sequence of points $\hat{s}_q \hat{s}_r \hat{s}_{max}$ is clockwise. Otherwise $\hat{s}_{max}$ is chosen such that the sequence of points $\hat{s}_q \hat{s}_r \hat{s}_{max}$ is

counter-clockwise (see following figure).

Reference

1.G. N. Ramachandran and S. Ramaeshan, "Crystal Optics," in v. XXV/1 of "*Handbuch der Physik*," ed. by S. Flugge (Springer-Verlag, 1961).

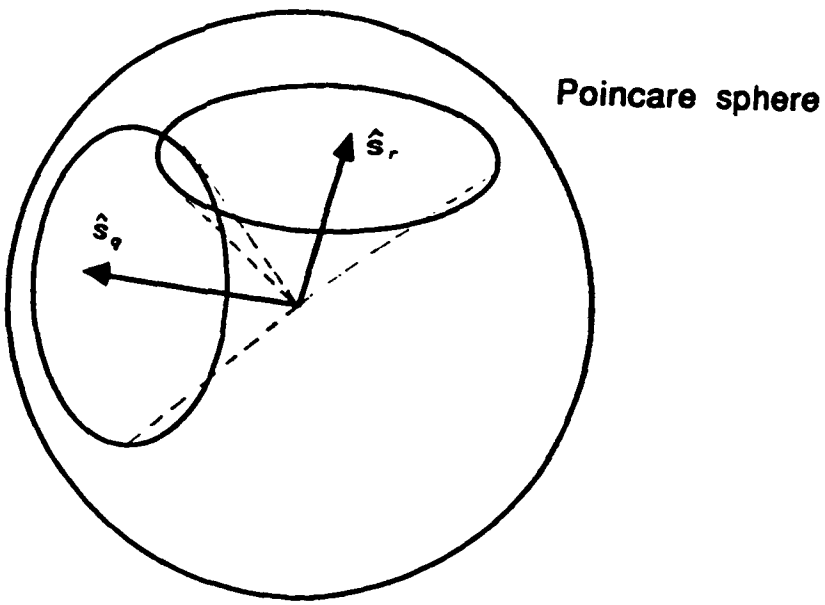


Fig. 1.

STUDIES ON MUELLER CALCULUS

Shih-Yau Lu
Russell A. Chipman
Department of Physics
University of Alabama in Huntsville

On relations between elements of the Mueller matrix

- a summary -

1. INTRODUCTION

Any 2x2 complex matrix is a valid Jones matrix. Similar statement is not true for the Mueller matrix. Hence there are constraints between elements of a Mueller matrix. For a non-deloparizing Mueller matrix, certain, in fact, nine, equalities must be satisfied by its elements. Since several papers have been written on this subject, in the following we mainly review the published results. The physically realizable problem is also discussed.

2. NON-DEPOLARIZING MUELLER MATRICES

In this section we consider only physically realizable Mueller matrix. It is known that a non-depolarizing Mueller matrix, say $\mathbb{M}$, can be derived from a Jones matrix, say J, by

$$\mathbb{M} = \mathbb{T}(J \otimes J^*)\mathbb{T}^{-1} \tag{1}$$

or

$$J \otimes J^* = \mathbb{T}^{-1}\mathbb{M}\mathbb{T} \tag{2}$$

where

$$\mathbb{T} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & i & -i & 0 \end{bmatrix} \tag{3}$$

From Eqs. (3) and (4) we see that $J_{kl}J^*_{mn}(k, l, m, n = 1, 2)$ is linear functions of M_{ij} $(i, j = 1, 2, 3, 4)$ matrix (Tab. 1), and vice verse (Tab. 2). A Jones matrix has eight degrees of freedom and a Mueller matrix has sixteen. Since a Mueller does not provide the information about total phase, nine equations must be satisfied by its elements.

A. Results of Refs. (2) and (3)

In general, any relation of the form

$$(J_{kl}J^*_{mn})(J_{pq}J^*_{rs}) = (J_{kl}J^*_{rs})(J_{pq}J^*_{mn}) \tag{4}$$

will yield a equation for elements of the Mueller matrix. Only nine equations are independent, and they are listed in Table 3.

The nine equations listed in Table 3 are not unique. Our choice coincides with Ref. 2 and 3. By summing first six equations in Table 3, it is obtained that[3,4,5]

$$\Sigma M_{ij} = 4M_{11} \tag{5}$$

Several authors[4-7] argued that Eq. (14) is the necessary and sufficient conditions for a Mueller matrix to be non-depolarizing.

B. Schaefer's results

The following equations[9] are directly resulted from equations in Table 3:

$$M_{11} + M_{12} \geq \pm(M_{31} + M_{32})$$
$$\geq \pm(M_{41} + M_{42}) \tag{6}$$
$$M_{11} - M_{12} \geq \pm(M_{31} - M_{32})$$
$$\geq \pm(M_{41} - M_{42}) \tag{7}$$
$$M_{11} + M_{21} \geq \pm(M_{13} + M_{23})$$

$$\geq \pm(M_{14} + M_{24}) \tag{8}$$

$$M_{11} - M_{21} \geq \pm(M_{13} - M_{23})$$

$$\geq \pm(M_{14} - M_{24}) \tag{9}$$

$$M_{11} + M_{22} \geq \pm(M_{33} + M_{44})$$

$$\geq \pm(M_{34} - M_{43}) \tag{10}$$

$$M_{11} - M_{22} \geq \pm(M_{33} - M_{44})$$

$$\geq \pm(M_{34} + M_{43}) \tag{11}$$

$$M_{11} \geq M_{ij}, \text{ for all } i, j. \tag{12}$$

Eq. (21) indicates that M_{11} is the maximum elements of a Mueller matrix.

C. Barakat's results

The input and output coherency matrices are related by

$$\bar{\mathbf{C}} = \mathbf{J}\mathbf{C}\mathbf{J}^{*}. \tag{13}$$

The coherency matrix C can be written as

$$\mathbf{C} = \begin{bmatrix} S_1 + S_2 & S_3 - iS_4 \\ S_3 + iS_4 & S_1 - S_2 \end{bmatrix} \tag{14}$$

By taking determinants of both sides of Eq. (13), we have

$$\bar{S}_1^2 - \bar{S}_2^2 - \bar{S}_3^2 - \bar{S}_4^2 = |Det\,\mathbf{J}|^2 (S_1^2 - S_2^2 - S_3^2 - S_4^2). \tag{15}$$

Let $\mathbf{G} = Diag(1, -1, -1, -1)$. Then Eq. (15) can be rewritten as

$$\bar{\mathbf{S}}^T \mathbf{G} \bar{\mathbf{S}} = |Det\,\mathbf{J}|^2 \mathbf{S}^T \mathbf{G} \mathbf{S} \tag{16}$$

where

$$\bar{\mathbf{S}} = \begin{bmatrix} \bar{S}_1 \\ \bar{S}_2 \\ \bar{S}_3 \\ \bar{S}_4 \end{bmatrix} \text{ and } \mathbf{S} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}.$$

Since we have

$$\bar{\mathbf{S}} = \mathbf{M}\,\mathbf{S}, \tag{17}$$

by Eq. (16) it is obtained that

$$\mathbf{M}^T \mathbf{G}\,\mathbf{M} = |Det\,\mathbf{J}|^2 \mathbf{G}. \tag{18}$$

Note that

$$|Det\,\mathbf{J}|^2 = M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 \tag{19}$$

From Eqs. (18) and (19), nine equations are found:

$$M_{11}M_{12} - M_{21}M_{22} - M_{31}M_{32} - M_{41}M_{42} = 0 \tag{20}$$

$$M_{11}M_{13} - M_{21}M_{23} - M_{31}M_{33} - M_{41}M_{43} = 0 \tag{21}$$

$$M_{11}M_{14} - M_{21}M_{24} - M_{31}M_{34} - M_{41}M_{44} = 0 \tag{22}$$

$$M_{12}M_{13} - M_{22}M_{23} - M_{32}M_{33} - M_{42}M_{43} = 0 \tag{23}$$

$$M_{12}M_{14} - M_{22}M_{24} - M_{32}M_{34} - M_{42}M_{44} = 0 \tag{24}$$

$$M_{13}M_{14} - M_{23}M_{24} - M_{33}M_{34} - M_{43}M_{44} = 0 \tag{25}$$

$$M_{12}^2 - M_{22}^2 - M_{32}^2 - M_{42}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0 \tag{26}$$

$$M_{13}^2 - M_{23}^2 - M_{33}^2 - M_{43}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0 \tag{27}$$

$$M_{14}^2 - M_{24}^2 - M_{34}^2 - M_{44}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0 \tag{28}$$

3. DEPOLARIZING MUELLER MATRICES

The depolarizing Mueller matrix is resulted from the average of different non-depolarizing Mueller matrices, for example scattering process. Another example is when light propagating in random medium. Thus Eqs. (1) and (2) can be rewritten as

$$\mathbf{M} = \mathbf{T} < \mathbf{J} \otimes \mathbf{J}^* > \mathbf{T}^{-1} \tag{29}$$

and

$$< \mathbf{J} \otimes \mathbf{J}^* > = \mathbf{T}^{-1} \mathbf{M} \mathbf{T}. \tag{30}$$

From Schwarz inequality, some of the equalities in Table 3 become inequalities[3]. These inequalities are listed in Table 4. It should be noted that these inequalities are the conditions satisfied by physically realizable Mueller matrices.

Let

$$\mathbf{N} = \sum_j \mathbf{M}_j \sigma_j \otimes \sigma_j . \tag{31}$$

Cloude[10] pointed that the necessary and sufficient conditions for a Mueller matrix to be physically realizable is that the eigenvalues of **N** are all nonnegative, i.e., **N** is positive semi-definite.

REFERENCES

1. E. L. O'Neil, *Introduction to Statical Optics* (Addison-Wesley, 1963).

2. K. D. Abhyankar and A. L. Fymat, "Relations between the elements of the phase matrix for scattering," J. Math. Phys. **10**, 1935-1938 (1969).

3. E. S. Fry and G. W. Kattawar, " Relationships between elements of the Stokes matrix," App. Opt. **20**, 2811-2814 (1981).

4. R. Simon, "The connection between Mueller and Jones matrices of polarization optics," Opt. Comm. **42**, 293-297 (1982).

5. K. Kim, L. Mandel and E. Wolf, "Relationship between Jones and Mueller matrices for random media," J. Opt. Soc. Am. **A4**, 433-437 (1987).

6. J. J. Gil, "A depolarization criterion in Mueller matrix," Optica Acta, **32**, 259-261 (1985).

7. R. Simon, "Mueller matrices and depolarization criteria," J. Mod. Opt. **43**, 569-575 (1987).

8. R. Barakat, "Blinear constraints between elements of the 4X4 Mueller-Jones transfer matrix of polarization theory," Opt. Comm. **38**, 159-161 (1981).

9. R. W. Schaefer, "Inequalities between the elements of the Mueller scattering matrix," App. Opt. **20**, 2875 (1981).

10. R. Cloude, " Conditions for the physically realizability of matrix operators in polarimetry," SPIE **1166**, 177-185 (1989).

TABLE 1

$$J_{11}J_{11}^{*} = \frac{1}{2}(M_{11} + M_{12} + M_{21} + M_{22}) \tag{J1}$$

$$J_{12}J_{12}^{*} = \frac{1}{2}(M_{11} - M_{12} + M_{21} - M_{22}) \tag{J2}$$

$$J_{21}J_{21}^{*} = \frac{1}{2}(M_{11} + M_{12} - M_{21} - M_{22}) \tag{J3}$$

$$J_{22}J_{22}^{*} = \frac{1}{2}(M_{11} - M_{12} - M_{21} + M_{22}) \tag{J4}$$

$$J_{11}J_{12}^{*} = \frac{1}{2}(M_{13} + M_{23} + iM_{14} + iM_{24}) \tag{J5}$$

$$J_{11}J_{21}^{*} = \frac{1}{2}(M_{31} + M_{32} - iM_{41} - iM_{42}) \tag{J6}$$

$$J_{11}J_{22}^{*} = \frac{1}{2}(M_{33} + M_{44} + iM_{34} - iM_{43}) \tag{J7}$$

$$J_{12}J_{21}^{*} = \frac{1}{2}(M_{33} - M_{44} - iM_{34} - iM_{43}) \tag{J8}$$

$$J_{12}J_{22}^{*} = \frac{1}{2}(M_{31} - M_{32} - iM_{41} + iM_{42}) \tag{J9}$$

$$J_{21}J_{22}^{*} = \frac{1}{2}(M_{13} - M_{23} + iM_{14} - iM_{24}) \tag{J10}$$

$$J_{22}J_{21}^{*} = (J_{21}J_{22}^{*})^{*} \tag{J11}$$

$$J_{22}J_{12}^{*} = (J_{12}J_{22}^{*})^{*} \tag{J12}$$

$$J_{21}J_{12}^{*} = (J_{12}J_{21}^{*})^{*} \tag{J13}$$

$$J_{22}J_{11}^{*} = (J_{11}J_{22}^{*})^{*} \tag{J14}$$

$$J_{21}J_{11}^{*} = (J_{11}J_{21}^{*})^{*} \tag{J15}$$

$$J_{12}J_{11}^{*} = (J_{11}J_{12}^{*})^{*} \tag{J16}$$

TABLE 2

$$M_{11} = \frac{1}{2}(J_{11}J_{11}^{*} + J_{12}J_{12}^{*} + J_{21}J_{21}^{*} + J_{22}J_{22}^{*}) \quad \text{(M1)}$$

$$M_{12} = \frac{1}{2}(J_{11}J_{11}^{*} - J_{12}J_{12}^{*} + J_{21}J_{21}^{*} - J_{22}J_{22}^{*}) \quad \text{(M2)}$$

$$M_{13} = \frac{1}{2}(J_{11}J_{12}^{*} + J_{12}J_{11}^{*} + J_{21}J_{22}^{*} + J_{22}J_{21}^{*}) \quad \text{(M3)}$$

$$M_{14} = \frac{1}{2}(-J_{11}J_{12}^{*} + J_{12}J_{11}^{*} - J_{21}J_{22}^{*} + J_{22}J_{21}^{*}) \quad \text{(M4)}$$

$$M_{21} = \frac{1}{2}(J_{11}J_{11}^{*} + J_{12}J_{12}^{*} - J_{21}J_{21}^{*} - J_{22}J_{22}^{*}) \quad \text{(M5)}$$

$$M_{22} = \frac{1}{2}(J_{11}J_{11}^{*} - J_{12}J_{12}^{*} - J_{21}J_{21}^{*} + J_{22}J_{22}^{*}) \quad \text{(M6)}$$

$$M_{23} = \frac{1}{2}(J_{11}J_{12}^{*} + J_{12}J_{11}^{*} - J_{21}J_{22}^{*} - J_{22}J_{21}^{*}) \quad \text{(M7)}$$

$$M_{24} = \frac{1}{2}(-J_{11}J_{12}^{*} + J_{12}J_{11}^{*} + J_{21}J_{22}^{*} - J_{22}J_{21}^{*}) \quad \text{(M8)}$$

$$M_{31} = \frac{1}{2}(J_{11}J_{21}^{*} + J_{21}J_{11}^{*} + J_{12}J_{22}^{*} + J_{22}J_{12}^{*}) \quad \text{(M9)}$$

$$M_{32} = \frac{1}{2}(J_{11}J_{21}^{*} + J_{21}J_{11}^{*} - J_{12}J_{22}^{*} - J_{22}J_{12}^{*}) \quad \text{(M10)}$$

$$M_{33} = \frac{1}{2}(J_{11}J_{22}^{*} + J_{22}J_{11}^{*} + J_{12}J_{21}^{*} + J_{21}J_{12}^{*}) \quad \text{(M11)}$$

$$M_{34} = \frac{1}{2}(-J_{11}J_{22}^{*} + J_{22}J_{11}^{*} + J_{12}J_{21}^{*} - J_{21}J_{12}^{*}) \quad \text{(M12)}$$

$$M_{41} = \frac{1}{2}(J_{11}J_{21}^{*} - J_{21}J_{11}^{*} + J_{12}J_{22}^{*} - J_{22}J_{12}^{*}) \quad \text{(M13)}$$

$$M_{42} = \frac{1}{2}(J_{11}J_{21}^{*} - J_{21}J_{11}^{2} - J_{12}J_{22}^{*} + J_{22}J_{12}^{*}) \quad \text{(M14)}$$

$$M_{43} = \frac{1}{2}(J_{11}J_{22}^{*} - J_{22}J_{11}^{*} + J_{12}J_{21}^{*} - J_{21}J_{12}^{*}) \quad \text{(M15)}$$

$$M_{44} = \frac{1}{2}(J_{11}J_{22}^{*} + J_{22}J_{11}^{*} - J_{12}J_{21}^{*} - J_{21}J_{12}^{*}) \quad \text{(M16)}$$

TABLE 3

$$4(J_{11}J_{11}^{*})(J_{21}J_{21}^{*}) = (M_{11}+M_{12})^2-(M_{21}+M_{22})^2$$
$$= (M_{31}+M_{32})^2+(M_{41}+M_{42})^2 = 4(J_{11}J_{21}^{*})(J_{21}J_{11}^{*})$$
$$4(J_{12}J_{12}^{*})(J_{22}J_{22}^{*}) = (M_{11}-M_{12})^2-(M_{21}-M_{22})^2$$
$$= (M_{31}-M_{32})^2+(M_{41}-M_{42})^2 = 4(J_{12}J_{22}^{*})(J_{22}J_{12}^{*})$$
$$4(J_{11}J_{11}^{*})(J_{12}J_{12}^{*}) = (M_{11}+M_{21})^2-(M_{12}+M_{22})^2$$
$$= (M_{13}+M_{23})^2+(M_{14}+M_{24})^2 = 4(J_{11}J_{12}^{*})(J_{12}J_{11}^{*})$$
$$4(J_{21}J_{21}^{*})(J_{22}J_{22}^{*}) = (M_{11}-M_{21})^2-(M_{12}-M_{22})^2$$
$$= (M_{13}-M_{23})^2+(M_{14}-M_{24})^2 = 4(J_{21}J_{22}^{*})(J_{22}J_{21}^{*})$$
$$4(J_{11}J_{11}^{*})(J_{22}J_{22}^{*}) = (M_{11}+M_{22})^2-(M_{12}+M_{21})^2$$
$$= (M_{33}+M_{44})^2+(M_{34}-M_{43})^2 = 4(J_{11}J_{22}^{*})(J_{22}J_{11}^{*})$$

$$4(J_{12}J_{12}^{*})(J_{21}J_{21}^{*}) = (M_{11}-M_{22})^2-(M_{12}-M_{21})^2$$
$$= (M_{33}-M_{44})^2+(M_{34}+M_{43})^2 = 4(J_{12}J_{21}^{*})(J_{21}J_{12}^{*})$$
$$4Re(J_{22}J_{11}^{*}J_{12}J_{21}^{*}) = M_{33}^2-M_{34}^2+M_{43}^2-M_{44}^2$$
$$= M_{13}^2-M_{14}^2-M_{23}^2+M_{24}^2 = 4Re(J_{22}J_{21}sper*J_{12}J_{11}^{*})$$
$$4Re(J_{22}J_{11}^{*}J_{21}J_{12}^{*}) = M_{33}^2-M_{43}^2+M_{34}^2-M_{44}^2$$

$$4Re(J_{22}J_{12}^{*}J_{11}J_{21}^{*}) = M_{31}^2-M_{32}^2+M_{41}^2-M_{42}^2$$
$$= M_{14}^2-M_{24}^2+M_{13}^2-M_{23}^2 = 4Re(J_{22}J_{21}^{*}J_{11}J_{12}^{*})$$

TABLE 4

$$4 < J_{11} J_{11}^{*} > < J_{21} J_{21}^{*} > = (M_{11} + M_{12})^2 - (M_{21} + M_{22})^2$$

$$\geq (M_{31} + M_{32})^2 + (M_{41} + M_{42})^2 = 4 | < J_{11} J_{21}^{*} > |^2$$

$$4 < J_{12} J_{12}^{*} > < J_{22} J_{22}^{*} > = (M_{11} - M_{12})^2 - (M_{21} - M_{22})^2$$

$$\geq (M_{31} - M_{32})^2 + (M_{41} - M_{42})^2 = 4 | < J_{12} J_{22}^{*} > |^2$$

$$4 < J_{11} J_{11}^{*} > < J_{12} J_{12}^{*} > = (M_{11} + M_{21})^2 - (M_{12} + M_{22})^2$$

$$\geq (M_{13} + M_{23})^2 + (M_{14} + M_{24})^2 = 4 | < J_{11} J_{12}^{*} > |^2$$

$$4 < J_{21} J_{21}^{*} > < J_{22} J_{22}^{*} > = (M_{11} - M_{21})^2 - (M_{12} - M_{22})^2$$

$$\geq (M_{13} - M_{23})^2 + (M_{14} - M_{24})^2 = 4 | < J_{21} J_{22}^{*} > |^2$$

$$4 < J_{11} J_{11}^{*} > < J_{22} J_{22}^{*} > = (M_{11} + M_{22})^2 - (M_{12} + M_{21})^2$$

$$\geq (M_{33} + M_{44})^2 + (M_{34} - M_{43})^2 = 4 | < J_{11} J_{22}^{*} > |^2$$

$$4 < J_{12} J_{12}^{*} > < J_{21} J_{21}^{*} > = (M_{11} - M_{22})^2 - (M_{12} - M_{21})^2$$

$$\geq (M_{33} - M_{44})^2 + (M_{34} + M_{43})^2 = 4 | < J_{12} J_{21}^{*} > |^2.$$

$$< J_{11} J_{11}^{*} > = M_{11} + M_{12} + M_{21} + M_{22} \geq 0$$

$$< J_{12} J_{12}^{*} > = M_{11} - M_{12} + M_{21} - M_{22} \geq 0$$

$$< J_{21} J_{21}^{*} > = M_{11} + M_{12} - M_{21} - M_{22} \geq 0$$

$$< J_{22} J_{22}^{*} > = M_{11} - M_{12} - M_{21} + M_{22} \geq 0$$

A simple derivation of the relations

between elements of non-depolarizing Mueller matrix

This note is to present a simpler derivation of Barakat's results on the relations batween elements of Mueller matrix. It is a simple physical fact that if the incident light of a non-deloparizing system is completely polarized, then so are the output light. In other words, $S_1^2 - S_2^2 - S_3^2 - S_4^2 = 0$ implies $\overline{S}_1^2 - \overline{S}_2^2 - \overline{S}_3^2 - \overline{S}_4^2 = 0$ where $S = (S_1, S_2, S_3, S_4)^T$ and $\overline{S} = (\overline{S}_1, \overline{S}_2, \overline{S}_3, \overline{S}_4)^T$ are the incident and output Stokes vectors, respectively. Thus the following equation is true:

$$(\overline{S}_1)^2 - (\overline{S}_2)^2 - (\overline{S}_3)^2 - (\overline{S}_4)^2 = \alpha(S_1^2 - S_2^2 - S_3^2 - S_4^2) \qquad (1)$$

where α is proportional constant depending only on the elements of the Mueller matrix. Let

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix}$$

be the Mueller matrix of the system under consideration. For unpolarized incident light, the output light is $\overline{S} = (M_{11}, M_{21}, M_{31}, M_{41})^T$, and then the α of Eq. (1) can be written as

$$\alpha = M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 . \qquad (2)$$

Then we can rewrite Eq. (1) as

$$(\overline{S}_1)^2 - (\overline{S}_2)^2 - (\overline{S}_3)^2 - (\overline{S}_4)^2$$

$$= (M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2)(S_1^2 - S_2^2 - S_3^2 - S_4^2). \tag{3}$$

Barakat derived Eq. (3) by the coherency matrix formalism. Our derivation which is only based upon a physical observation, is more straightforward.

Since $\mathbf{S}$ and $\bar{\mathbf{S}}$ are related by

$$\bar{\mathbf{S}} = \mathbf{M}\mathbf{S}, \tag{4}$$

from Eq. (3) we have

$$(\Sigma M_{1i} S_i)^2 - (\Sigma M_{2j} S_j)^2 - (\Sigma M_{3k} S_k)^2 - (\Sigma M_{4l} S_l)^2$$
$$= (M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2)(S_1^2 - S_2^2 - S_3^2 - S_4^2). \tag{5}$$

Compare the coefficients of both sides of Eq. (5), nine relations that are first obtained by Barakat, then, can be found:

$$M_{11}M_{12} - M_{21}M_{22} - M_{31}M_{32} - M_{41}M_{42} = 0$$
$$M_{11}M_{13} - M_{21}M_{23} - M_{31}M_{33} - M_{41}M_{43} = 0$$
$$M_{11}M_{14} - M_{21}M_{24} - M_{31}M_{34} - M_{41}M_{44} = 0$$
$$M_{12}M_{13} - M_{22}M_{23} - M_{32}M_{33} - M_{42}M_{43} = 0$$
$$M_{12}M_{14} - M_{22}M_{24} - M_{32}M_{34} - M_{42}M_{44} = 0$$
$$M_{13}M_{14} - M_{23}M_{24} - M_{33}M_{34} - M_{43}M_{44} = 0$$
$$M_{12}^2 - M_{22}^2 - M_{32}^2 - M_{42}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0$$
$$M_{13}^2 - M_{23}^2 - M_{33}^2 - M_{43}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0$$
$$M_{14}^2 - M_{24}^2 - M_{34}^2 - M_{44}^2 + M_{11}^2 - M_{21}^2 - M_{31}^2 - M_{41}^2 = 0$$

REFERENCE

R. Barakat, "Bilinear constraints between elements of the 4x4 Mueller-Jones transfer matrix of polarization theory," Opt. Commun. **38**, 159-161 (1981).

A study of the Mueller matrix in exponential from

1. A PARTITION OF THE MUELLER MATRIX

Let the Mueller matrix under consideration be denoted by

$$\mathbf{M} = \left[\begin{array}{c|ccc} M_{11} & M_{12} & M_{13} & M_{14} \\ \hline M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{array}\right] = \begin{bmatrix} M_{11} & \mathbf{A}^T \\ \mathbf{B} & \mathbf{C} \end{bmatrix} \tag{1}$$

where

$$\mathbf{A} = \begin{bmatrix} M_{12} \\ M_{13} \\ M_{14} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} M_{21} \\ M_{31} \\ M_{41} \end{bmatrix} \text{and } \mathbf{C} = \begin{bmatrix} M_{22} & M_{23} & M_{24} \\ M_{32} & M_{33} & M_{34} \\ M_{42} & M_{43} & M_{44} \end{bmatrix}.$$

The partition of a Mueller matrix in Eq. (1) will be proved to be useful. For example, under a rotation in Poincare sphere, $\mathbf{M}$ becomes

$$\overline{\mathbf{M}} = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{R} \end{bmatrix} \begin{bmatrix} M_{11} & \mathbf{A}^T \\ \mathbf{B} & \mathbf{C} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{R}^T \end{bmatrix} = \begin{bmatrix} M_{11} & (\mathbf{RA})^T \\ \mathbf{RB} & \mathbf{RCR}^T \end{bmatrix} \tag{2}$$

where $\mathbf{R}$ is a 3-dimensional rotation matrix, and

$$\mathbf{R}^T = \mathbf{R}^{-1}. \tag{3}$$

Besides, the equation

$$\mathbf{S}' = \mathbf{M}\mathbf{S} \tag{4}$$

can be written as

$$S_1' = M_{11}S_1 + \mathbf{A}^T\vec{S}, \text{ and} \tag{5}$$

$$\vec{\mathbf{S}}' = \mathbf{B}S_1 + \mathbf{C}\vec{S} \tag{6}$$

in which

$$\mathbf{S} = \begin{bmatrix} S_1 \\ \vec{S} \end{bmatrix} \text{ and } \mathbf{S}' = \begin{bmatrix} S_1' \\ \vec{S}' \end{bmatrix}.$$

In next section, we will use above partition to discuss the exponential form of the Mueller matrix.

2. THE EXPONENTIAL FORM

It was pointed out by Azzam that the Stokes vector of light propagating in continuous medium, homogeneous or not, can be described by a differential equation:

$$\frac{d\mathbf{S}}{dz} = \mathbf{m}\mathbf{S}. \tag{7}$$

In case that $\mathbf{m}$ is independent of z, the general solution of Eq. (7) is

$$\mathbf{S}(z) = e^{\mathbf{m}z}\mathbf{S}(0), \tag{8}$$

and we have

$$\mathbf{M}(z) = e^{\mathbf{m}z} \tag{9}$$

such that

$$\mathbf{S}(z) = \mathbf{M}(z)\mathbf{S}(0). \tag{10}$$

Thus $\mathbf{m}$ can be regarded as a exponential representation of the Mueller matrix.

The Stokes vector satisfies Eq. (7), but we want to know equation satisfied by the degree of polarization. First, we partition $\mathbf{m}$ into

$$\mathbf{m} = \begin{bmatrix} m_{11} & \mathbf{a}^T \\ \mathbf{b} & \mathbf{c} \end{bmatrix} \tag{11}$$

by the same way as of Eq. (1). Then Eqs. (7) can be rewritten as

$$\frac{dS_1}{dz} = m_{11} S_1 + \mathbf{a}^T \vec{S}, \text{ and} \tag{12}$$

$$\frac{d\vec{S}}{dz} = \mathbf{b} S_1 + \mathbf{c} \vec{S}. \tag{13}$$

And we have

$$\frac{d}{dz} P^2 = \frac{d}{dz}\left(\frac{\vec{S}^2}{S_1^2}\right)$$

and

$$\frac{d}{dz} P^2 = 2\left\{\frac{\vec{S}^T(\mathbf{c} - m_{11})\vec{S}}{S_1^2} + \frac{\mathbf{b}^T\vec{S}}{S_1} - \frac{(\vec{S}^T\vec{S})\mathbf{a}^T\vec{S}}{S_1^3}\right\}, \tag{14}$$

in which P denotes the degree of polarization (DOP). Eq. (14) is the equation governed the square of DOP.

3. NON-DEPOLARIZING CASES

For non-depolarizing medium, completely polarized light can only produce completely polarized light, i.e., $dP^2/dz = 0$ if $\vec{S}^T\vec{S} = S_1^2$. In this case Eq. (14) can be written as

$$\frac{d}{dz}P^2 = 2\left\{\frac{\bar{S}^T(\mathbf{c}-m_{11})\bar{S}}{S_1^2} + \frac{(\mathbf{b}^T-\mathbf{a}^T)\bar{S}}{S_1}\right\}, \tag{15}$$

or

$$\frac{d}{dz}P^2 = 2\{\hat{S}^T(\mathbf{c}-m_{11})\hat{S} + (\mathbf{b}-\mathbf{a})^T\hat{S}\} \tag{16}$$

in which $\hat{S}$ denotes the unit vector $\begin{bmatrix} S_2/S_1 \\ S_3/S_1 \\ S_4/S_1 \end{bmatrix}$. Eq. (15) or (16) equal to zero, only if

$\mathbf{c} - m_{11}$ is anti-symmetric and $\mathbf{a} = \mathbf{b}$. Therefore, for non-depolarization medium, the general form of $\mathbf{m}$ is

$$\mathbf{m} = \begin{bmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \mu & \nu \\ \gamma & -\mu & \alpha & \eta \\ \delta & -\nu & -\eta & \alpha \end{bmatrix}. \tag{17}$$

This is exactly the results obtained by Azzam in a paper in 1978. It is quite clear, as pointed out by Azzam, that the meaning of parameters in Eq. (17) can be explained as follows:

α	total absorption
β	linear diattenuation
γ	linear diattenuation at 45°
δ	circular diattenuation

η	linear retardance
ν	linear retardance at 45°
μ	circular retardance

REFERENCE

R. M. A. Azzam, "Propagation of partially polarized light through anisotropic media with or without depolarization: A differential 4x4 matrix calculus," JOSA, **68**, 1756-1767 (1978).

On the relations between the elements

of a Mueller matrix

Consider an arbitrary non-depolarizing system. Its Mueller matrix and corresponding Jones Matrix are denoted by $\mathbf{M} = [M_{ij}]$ and $\mathbf{J} = [J_{mn}]$. It is known that

$$|J_{11}|^2 = \tfrac{1}{2}(M_{11} + M_{12} + M_{21} + M_{22}) \geq 0 \tag{1}$$

$$|J_{12}|^2 = \tfrac{1}{2}(M_{11} - M_{12} + M_{21} - M_{22}) \geq 0 \tag{2}$$

$$|J_{21}|^2 = \tfrac{1}{2}(M_{11} + M_{12} - M_{21} - M_{22}) \geq 0 \tag{3}$$

$$|J_{22}|^2 = \tfrac{1}{2}(M_{11} - M_{12} - M_{21} + M_{22}) \geq 0 . \tag{4}$$

The product of physically realizable Mueller matrices is a physically realizable Mueller matrix. Moreover, the product of non-depolarizing Mueller matrices is a non-depolarizing Mueller matrix. The Mueller and Jones matrices of a left handed QWCR are, respectively,

$$\mathbf{M}_c = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{J}_c = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

So the following Mueller matrix $\mathbf{M}'$ are non-depolarizing and its Jones matrix is denoted by $\mathbf{J}'$:

$$\mathbf{M}' = \mathbf{M}_c \mathbf{M} \mathbf{M}_c^T = \begin{bmatrix} M_{11} & M_{13} & -M_{12} & M_{14} \\ M_{31} & M_{33} & -M_{32} & M_{34} \\ -M_{21} & -M_{23} & M_{22} & -M_{24} \\ M_{41} & M_{43} & -M_{42} & M_{44} \end{bmatrix}$$

$$\mathbf{J}' = \mathbf{J}_c \mathbf{J} \mathbf{J}_c^{-1} = \frac{1}{2}\begin{bmatrix} J_{11} + J_{12} + J_{21} + J_{22} & -J_{11} + J_{12} - J_{21} + J_{22} \\ -J_{11} - J_{12} + J_{21} + J_{22} & J_{11} - J_{12} - J_{21} + J_{22} \end{bmatrix}.$$

And then we have

$$\frac{1}{4}|J_{11} + J_{12} + J_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} + M_{13} + M_{31} + M_{33}) \geq 0 \quad (5)$$

$$\frac{1}{4}|-J_{11} + J_{12} - J_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} - M_{13} + M_{31} - M_{33}) \geq 0 \quad (6)$$

$$\frac{1}{4}|-J_{11} - J_{12} + J_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} + M_{13} - M_{31} - M_{33}) \geq 0 \quad (7)$$

$$\frac{1}{4}|J_{11} - J_{12} - J_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} - M_{13} - M_{31} + M_{33}) \geq 0 . \quad (8)$$

The Mueller and Jones matrices of a QWLR at 135° are

$$\mathbf{M}_l = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$

$$\mathbf{J}_l = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}.$$

The following Mueller matrix $\mathbf{M}''$ are non-depolarizing and its Jones matrix is denoted by $\mathbf{J}''$:

$$\mathbf{M}'' = \mathbf{M}_l \mathbf{M} \mathbf{M}_l^T = \begin{bmatrix} M_{11} & M_{14} & M_{13} & -M_{12} \\ M_{41} & M_{44} & M_{43} & -M_{42} \\ M_{31} & M_{34} & M_{33} & -M_{32} \\ -M_{21} & -M_{24} & -M_{23} & M_{22} \end{bmatrix}$$

$$\mathbf{J}'' = \mathbf{J}_l \mathbf{J} \mathbf{J}_l^{-1} = \frac{1}{2}\begin{bmatrix} J_{11} + iJ_{12} - iJ_{21} + J_{22} & iJ_{11} + J_{12} + J_{21} - iJ_{22} \\ -iJ_{11} + J_{12} + J_{21} + iJ_{22} & J_{11} - iJ_{12} + iJ_{21} + J_{22} \end{bmatrix}.$$

And then we have

$$\frac{1}{4}|J_{11} + iJ_{12} - iJ_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} + M_{14} + M_{41} + M_{44}) \geq 0 \quad (9)$$

$$\frac{1}{4}|iJ_{11} + J_{12} + J_{21} - iJ_{22}|^2 = \frac{1}{2}(M_{11} - M_{14} + M_{41} - M_{44}) \geq 0 \quad (10)$$

$$\frac{1}{4}|-iJ_{11} + J_{12} + J_{21} + iJ_{22}|^2 = \frac{1}{2}(M_{11} + M_{14} - M_{41} - M_{44}) \geq 0 \quad (11)$$

$$\frac{1}{4}|J_{11} - iJ_{12} + iJ_{21} + J_{22}|^2 = \frac{1}{2}(M_{11} - M_{14} - M_{41} + M_{44}) \geq 0 . \quad (12)$$

The Eqs. (5-8) or (9-12) can be regarded as a generalization of Eqs. (1-4).

On the non-depolarizing Mueller matrices

In this note, a statement in the book by Azzam and Bashara will be proved to be incorrect and, therefore, clarify the concept of non-depolarizing Mueller matrix.

The following words are copied form the book by Azzam and Bashara:

> **Therefore, for a non-depolarizing optical system, the degree of polarization [§1.8] of the output light from the system $\mathcal{P}_o$ is either greater than or equal to the degree of polarization of the input light $\mathcal{P}_i$, or,**
>
> $$\mathcal{P}_o \geq \mathcal{P}_i, \qquad (2.222a)$$
>
> **for *all* incident states of (total or partial) polarization. The case of depolarizing optical systems that include incoherent scattering processes and for which**
>
> $$\mathcal{P}_o \leq \mathcal{P}_i, \qquad (2.222b)$$
>
> **at least for *one* incident state will be studied in §2.12.**

But some non-depolarizing systems indeed can have 'depolarization effect' for some partially polarized states. For example, consider a non-depolarizing system, in fact a linear diattenuator, with a Jones matrix

$$J = \begin{bmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix} \qquad (1)$$

When the incident light is polarized light with a coherency matrix

$$C_i = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \qquad (2)$$

the coherency matrix of the output light is then

$$C_o = J C_i J^*$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this example, the output light can be unpolarized, even though the incident light is not. Therefore, $P_o \geq P_i$ is not always true for non-depolarizing system. This example indicates that a non-depolarizing system may decrease the DOP for partially polarized incident light. The statement quoted above is not true.

The definition of non-depolarizing Mueller matrix is that

(a) this Mueller matrix can be derived from a Jones matrix, or

(b) completely polarized light produces only completely polarized light.

Definition (b) is a simple but powerful statement. From it all the properties of the non-depolarizing Mueller matrix can be derived. On the other hand, the definition of depolarizing Mueller matrix is that completely polarized light may produce partially polarized light. It should be noted that a non-depolarizing Mueller matrix does not decrease the DOP for completely polarized light, but a depolarizing Mueller matrix does not increase the DOP for completely polarized light. In other words, Azzam and Bashara's Eqs. (2.222a) and (2.222b) is true only for completely polarized incident light, not for all incident states of (total or partial) polarization.

REFERENCE

R. M. A. Azzam and N. M. Bashara, *Ellipsometry and Polarized Light*, pp.141 (North-Holland, 1977).

THE INTERPRETATION OF MUELLER MATRICES

Shih-Yau Lu
Russell A. Chipman
Department of Physics
University of Alabama in Huntsvills

Dec 9, 1991

Statement of Work

Mueller matrices are now being routinely measured.

Their interpretation in terms of the associated polarization phenomena are not yet clear.

Also, how do we interpret inhomogeneous polarization elements?

Outline of Research Program

1. Analyze inhomogeneous polarization elements within the Jones calculus

2. Analyze the exponential form of Jones matrices

3. Establish and publish data reduction algorithms for Jones matrices

4. Map these results into the non-depolarizing subspace of the Mueller matrices

5. Establish algorithms for data reduction of non-depolarizing Mueller matrices

6. Establish and publish data reduction algorithms for Mueller calculus

Reasons for Interest in Inhomogeneous Polarization Elements

1. Combinations of polarization elements are usually inhomogeneous

2. Skew rays through optical systems are slightly inhomogeneous.

3. It is necessary to separate the inhomogeneous part of the Mueller matrix from the depolarizing part in data reduction

Properties of Inhomogeneous Polarization Elements

1. Established a measure of inhomogeneity

   $\eta = |\hat{E}_q^{+}\hat{E}_r|, 0 \leq \eta < 1.$

   $\hat{E}_q$ and $\hat{E}_r$ are eigenpolarization.

2. η relates to the nonorthogonality of eigenpolarizations.

3. T_{max} and T_{min} are associated with orthogonal polarization states

4. Polarization behavior is well characterized by eigenvalues and η.

5. Definitions of diattenuation and retardance for inhomogeneous polarization elements

Homogeneous and Inhomogeneous Polarization Elements

Eigenpolarizations

$$J\hat{E}_q = \xi_q \hat{E}_q \ , \ J\hat{E}_r = \xi_r \hat{E}_r$$

Homogeneous polarization elements

Orthogonal eigenpolarizations

$$\hat{E}_q^+ \hat{E}_r = 0$$

Diattenuation: $\mathcal{D} = \dfrac{\left| |\xi_q|^2 - |\xi_r|^2 \right|}{|\xi_q|^2 + |\xi_r|^2} \ , \quad 0 \le \mathcal{D} \le 1$

Retardance: $\mathcal{R} = \left| \delta_q - \delta_r \right| \ , \quad 0 \le \mathcal{R} \le \pi$

Note:

$|\xi_q|^2, |\xi_r|^2$ **are** $T_{\max}, T_{\min}$

$\hat{E}_q, \hat{E}_r$ **are** $\hat{E}_{\max}, \hat{E}_{\min}$

Inhomogeneous polarization elements

Nonorthogonal eigenpolarizations

$$\hat{E}_q^+ \hat{E}_r \neq 0$$

What are diattenuation and retardance?

Let $\left|\hat{E}_q^{+}\hat{E}_r\right| = \eta$

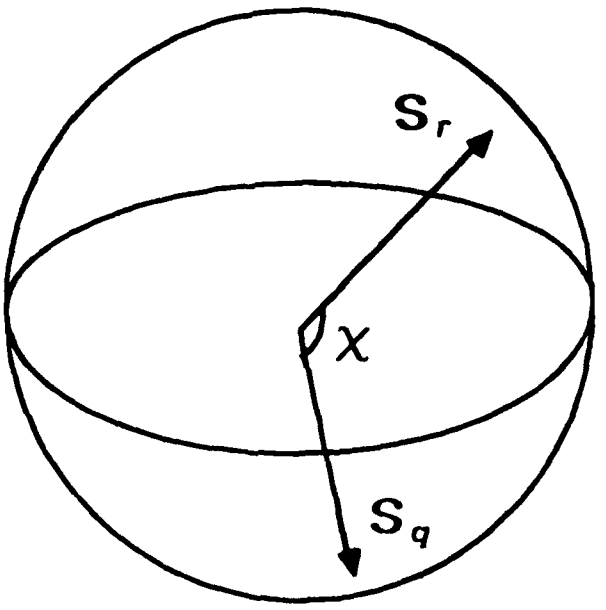
$0 \leq \eta < 1$

$\eta = 0$ **for homogeneous polarization elements**

$\eta =$ **inhomogeneity**

$\eta = \cos\frac{\chi}{2}$ **or** $\chi = 2\cos^{-1}\eta$ **,** χ **is the angle between eigenpolarizations on the Poincare sphere**

Poincare sphere



Polar Decomposition of Jones Matrices

Any polarization element is equivalent to a cascade of homogeneous diattenuator and homogeneous retarder.

$$\mathbf{J} = \mathbf{J}_R \mathbf{J}_D = \mathbf{J}_D{}' \mathbf{J}_R$$

$\mathbf{J}_D$ **and** $\mathbf{J}_D{}'$ **are uniquely defined**

In general, $\mathbf{J}_D \neq \mathbf{J}_D{}'$**, but they have same eigenvalues**

$\mathbf{J}_R$ **is unique only when** $Det\,\mathbf{J} \neq 0$

$$\mathcal{D}(\mathbf{J}) = \mathcal{D}(\mathbf{J}_D) = \mathcal{D}(\mathbf{J}_D{}')$$

$$\mathcal{R}(\mathbf{J}) = \mathcal{R}(\mathbf{J}_R)$$

Eigenvalues of J_D (or J_D) are $\sqrt{T_{max}}, \sqrt{T_{min}}$

$$\hat{E}^{+}_{max}\hat{E}_{min} = 0$$

$$(J\hat{E}_{max})^{+}(J\hat{E}_{min}) = 0$$

$$\mathcal{D} = \frac{T_{max} - T_{min}}{T_{max} + T_{min}}$$

D is equal to the DOP of the output light when incident light is unpolarized, i.e.,

$$D = DOP(U)$$

$$U = \hat{E}_{max} \oplus \hat{E}_{min}$$

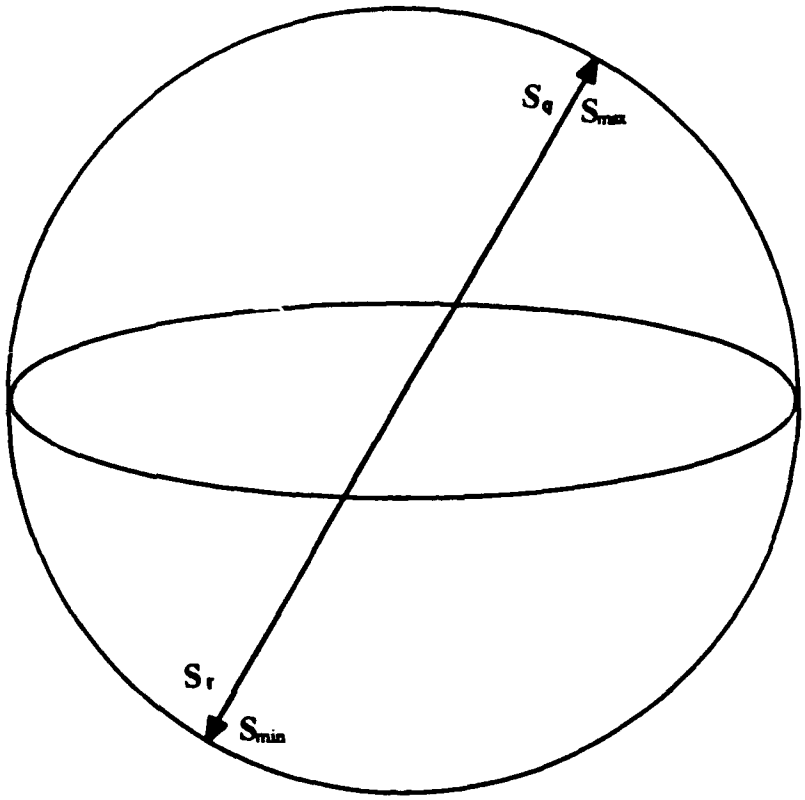
$$JU = J\hat{E}_{max} \oplus J\hat{E}_{min}$$

If $Det\, J = 0$,

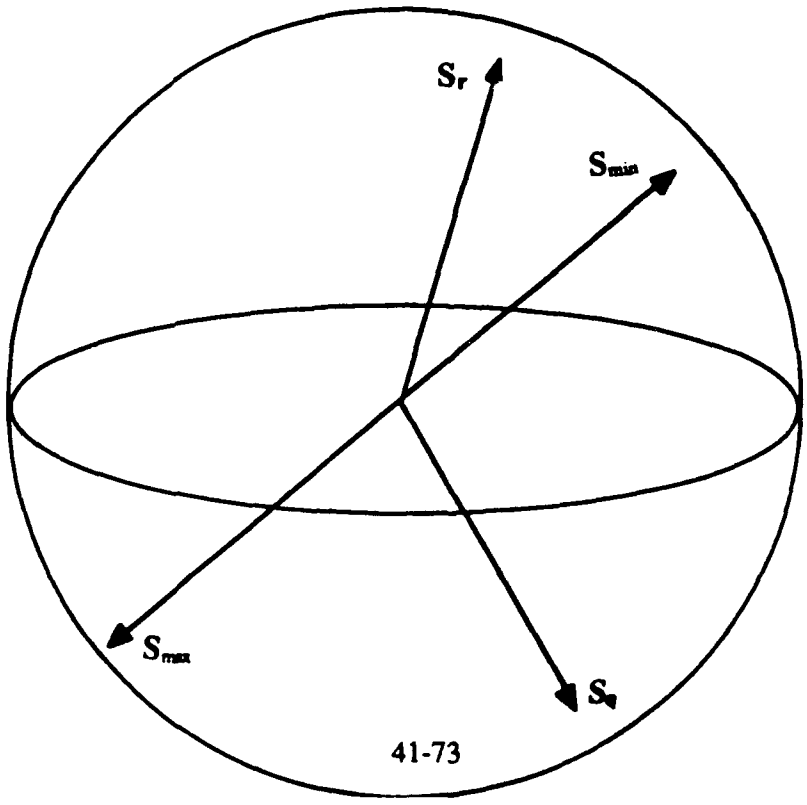
$$\mathcal{D} = 1, \ \mathcal{R} = 0$$

polarizer- output light is always at a fixed polarization state

Poincare Sphere: Homogeneous



Poincare Sphere: Inhomogeneous



Inhomogeneous Polarization Elements

T_{max} **and** T_{min} **are extreme values of** $T = |\mathbf{J}\hat{\mathbf{E}}|^2$

$$T_{max} T_{min} = |\xi_q|^2 |\xi_r|^2 = |Det\,\mathbf{J}|^2$$

$$T_{max} + T_{min} = \frac{|\xi_q|^2 + |\xi_r|^2 - \eta^2\left(\xi_q \xi_r^* + \xi_q^* \xi_r\right)}{1-\eta^2}$$

Thus $\mathcal{D}$ **and** $\mathcal{R}$ **are given by**

$$\mathcal{D} = \left\{ 1 - \frac{4(1-\eta^2)^2 |\xi_q|^2 |\xi_r|^2}{\left[|\xi_q|^2 + |\xi_r|^2 - \eta^2\left(\xi_q \xi_r^* + \xi_q^* \xi_r\right)\right]^2} \right\}^{1/2}$$

$$\mathcal{R} = 2\cos^{-1}\left\{ \left[\frac{(1-\eta^2)(|\xi_q| + |\xi_r|)^2}{(|\xi_q| + |\xi_r|)^2 - \eta^2\left(2|\xi_q||\xi_r| + \xi_q \xi_r^* + \xi_q^* \xi_r\right)} \right]^{1/2} \left| \cos\frac{\delta_q - \delta_r}{2} \right| \right\}$$

Degenerate polarization elements

Only one eigenpolarization $\hat{E}_q$

Example: $LP(0°)HWLR(45°) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

$$= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

Redefine $|(J - \xi_q I)\hat{F}| = \eta$ **where** $\hat{F}$ **is normal to the eigenpolarization** $\hat{E}_q$.

$$\mathcal{D} = \frac{\eta}{2|\xi_q|^2 + \eta^2}\sqrt{4|\xi_q|^2 + \eta^2}$$

$$\mathcal{R} = 2\cos^{-1}\frac{|\xi_q|}{\sqrt{|\xi_q|^2 + \frac{\eta^2}{4}}}$$

There exists a unitary matrix U **such that**

$$UJU^{+} = \begin{bmatrix} \xi_q & \eta \\ 0 & \xi_q \end{bmatrix}$$

An Unitarily Equivalent Theorem and Classification of Polarization Elements

Any two Jones matrices with the same eigenvalues and η are unitarily equivalent, i.e., there exist an unitary matrix U such that

$$U J_1 U^{+} = J_2 .$$

$PE(\xi_q, \xi_r; \eta)$ can be used to denote the set of unitarily equivalent polarization elements with the same eigenvalues and η.

Exponential Representation of Jones Matrices

Examples of decomposition

(1)

$$J = (\alpha_0\sigma_0 + \alpha_1\sigma_1 + \alpha_2\sigma_2 + \alpha_3\sigma_3)\exp i(\delta_0\sigma_0 + \delta_1\sigma_1 + \delta_2\sigma_2 + \delta_3\sigma_3)$$

(2) $J = c_0\sigma_0 + c_1\sigma_1 + c_2\sigma_2 + c_3\sigma_3$

(3) $J = \exp i(d_0\sigma_0 + d_1\sigma_1 + d_2\sigma_2 + d_3\sigma_3)$

What is the best form for interpreting the properties (i.e., diattenuation, retardance, inhomogenity) of polarization elements?

(1) has its use

(2) is better for routine manipulation

(3) is better to address deeper questions of the fundamental properties and the structure of the calculus.

An identity

$$\lim_{N\to\infty}\left(\exp\frac{J_1}{N}\exp\frac{J_2}{N}\right)^N$$

$$=\lim_{N\to\infty}\left(\exp\frac{J_2}{N}\exp\frac{J_1}{N}\right)^N$$

$$=\exp(J_1+J_2)$$

Generalization

$$\lim_{N\to\infty}\left(\exp\frac{d_0\sigma_0}{N}\exp\frac{d_1\sigma_1}{N}\exp\frac{d_2\sigma_2}{N}\exp\frac{d_3\sigma_3}{N}\right)^N$$

$$=\exp(d_0\sigma_0+d_1\sigma_1+d_2\sigma_2+d_3\sigma_3)$$

Similarly

$$\lim_{N\to\infty}\exp\frac{\Re d_0\sigma_0+\Re d_1\sigma_1+\Re d_2\sigma_2+\Re d_3\sigma_3}{N}\exp i\frac{\Im d_0\sigma_0+\Im d_1\sigma_1+\Im d_2\sigma_2+\Im d_3\sigma_3}{N})^N$$

$$=\exp(d_0\sigma_0+d_1\sigma_1+d_2\sigma_2+d_3\sigma_3)$$

$\Re d_0$ = **real part of** d_0

$\Im d_0$ = **imaginary part of** d_0

etc.

Canonical Forms

Total amplitude and phase

$$\exp(d_0\sigma_0)$$

Linear polarization element at $0°$

$$\exp(d_1\sigma_1)$$

$$\Delta\phi = 2\Im d_1, D = \tanh 2\Re d_1$$

Linear polarization element at $45°$

$$\exp(d_2\sigma_2)$$

$$\Delta\phi = 2\Im d_2, D = \tanh 2\Re d_2$$

Circular polarization element

$$\exp(d_3\sigma_3)$$

$$\Delta\phi = 2\Im d_3, D = \tanh 2\Re d_3$$

Transforming Jones Matrices in Pauli Spin Matrix Form to Mueller Matrices

Jones Matrix

$$J = \rho_0\sigma_o + \rho_1 e^{i\delta_1}\sigma_1 + \rho_2 e^{i\delta_2}\sigma_2 + \rho_3 e^{i\delta_3}\sigma_3$$

Equivalent Mueller Matrix

$$M = \begin{pmatrix} \rho_0^2+\rho_1^2+\rho_2^2+\rho_3^2 & 2\rho_0\rho_1\cos\delta_1 + 2\rho_2\rho_3\sin(\delta_2-\delta_3) & 2\rho_0\rho_2\cos\delta_2 + 2\rho_1\rho_3\sin(\delta_3-\delta_1) & 2\rho_0\rho_3\cos\delta_3 + 2\rho_1\rho_2\sin(\delta_1-\delta_2) \\ 2\rho_0\rho_1\cos\delta_1 - 2\rho_2\rho_3\sin(\delta_2-\delta_3) & \rho_0^2+\rho_1^2-\rho_2^2-\rho_3^2 & 2\rho_0\rho_3\sin\delta_3 + 2\rho_1\rho_2\cos(\delta_1-\delta_2) & -2\rho_0\rho_2\sin\delta_2 + 2\rho_1\rho_3\cos(\delta_3-\delta_1) \\ 2\rho_0\rho_2\cos\delta_2 - 2\rho_1\rho_3\sin(\delta_3-\delta_1) & -2\rho_0\rho_3\sin\delta_3 + 2\rho_1\rho_2\cos(\delta_1-\delta_2) & \rho_0^2-\rho_1^2+\rho_2^2-\rho_3^2 & 2\rho_0\rho_1\sin\delta_1 + 2\rho_2\rho_3\cos(\delta_2-\delta_3) \\ 2\rho_0\rho_3\cos\delta_3 - 2\rho_1\rho_2\sin(\delta_1-\delta_2) & 2\rho_0\rho_2\sin\delta_2 + 2\rho_1\rho_3\cos(\delta_3-\delta_1) & -2\rho_0\rho_1\sin\delta_1 + 2\rho_2\rho_3\cos(\delta_2-\delta_3) & \rho_0^2-\rho_1^2-\rho_2^2+\rho_3^2 \end{pmatrix}$$

COMPARISON OF THE JONES AND MUELLER CALCULUS

	Jones	Mueller
Type	Amplitude	Intensity
Polarization State	Jones Vector 2 complex elements 4 parameters	Stokes Vector 4 real elements 4 parameters
Describes Absolute Phase	Yes	No
Describes Partial Polarized Light	No	Yes
Polarization Elements	Jones Matrix 2 x 2 complex elements 8 parameters	Mueller Matrix 4 x 4 real elements 16 parameters
Describes Phase Change	Yes	No
Describes Depolarization	No	Yes
Ease of Use	Simpler	Involved
Strengths	Theory	Experimental
Applications Experimental	Interferometry	Polarimetry Scattering
Theoretical	Interferometry Diffraction Theory Aberration Theory Polarimetry	Polarimetry Scattering Theory Radiative Transfer

Depolarization

The coupling of polarized light into unpolarized light.

Diattenuation and retardance which vary rapidly in space and/or time.

Intrinsically associated with scattering.

Milk	Ground Glass
Clouds	Optical Fibers
Rough Surfaces	

Ideal Depolarizer

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} s_0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Partial Depolarizer

Incident Polarized light is reduced to a degree of polarization, d

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & d & 0 & 0 \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & d \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = (1-d) \begin{pmatrix} s_0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + d \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix}$$

Reasons for interest in depolarization

1. Loss of coherence is an irrecoverable loss of information in optical processors and optical computers.

2. Depolarization conveys information on the statistics of optical media and interfaces.

3. Depolarization may be a useful tool for investigating thin films microstructure.

4. Depolarization may limit performeance of polarizers and organic thin films.

5. Depolarization of laser radar targets is an area in need of development.

How many types of depolarization are there?

7 degrees of freedom in non-depolarizing Mueller matrix (8 elements in Jones matix minus absolute phase)

16 degrees of freedom in Mueller matrix

9 potential degrees of freedom for depolarization

Do the inequalities between elements of the Mueller matrix reduce the number of degrees of freedom? Probably not.

A Brief Analysis of Depolarizing Mueller Matrices

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix}$$

For maximum transmittance :

$$T_{max} = m_{11} + \sqrt{m_{12}^2 + m_{13}^2 + m_{14}^2}$$

$$\hat{S}_{max} = \begin{bmatrix} 1 \\ \dfrac{m_{12}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \\ \dfrac{m_{13}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \\ \dfrac{m_{14}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \end{bmatrix}$$

For minimum transmittance :

$$T_{min} = m_{11} - \sqrt{m_{12}^2 + m_{13}^2 + m_{14}^2}$$

$$\hat{S}_{min} = \begin{bmatrix} 1 \\ \dfrac{-m_{12}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \\ \dfrac{-m_{13}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \\ \dfrac{-m_{14}}{\sqrt{m_{12}^2+m_{13}^2+m_{14}^2}} \end{bmatrix}$$

Define

Diattenuation $\mathcal{D} = \frac{T_{max} - T_{min}}{T_{max} + T_{min}} = \frac{\sqrt{m_{12}^2 + m_{13}^2 + m_{14}^2}}{m_{11}}$

Polarizance $\mathcal{P} = DOP(\mathbf{U}) = \frac{\sqrt{m_{21}^2 + m_{31}^2 + m_{41}^2}}{m_{11}}$

$\hat{\mathbf{S}}_{max} \perp \hat{\mathbf{S}}_{min}$

in general, $\mathbf{M}\hat{\mathbf{S}}_{max} \not\perp \mathbf{M}\hat{\mathbf{S}}_{min}$

$\mathcal{D} \neq \mathcal{P}$

How to define $\mathcal{R}$**?**

Depolarizing index = $\frac{1}{4\pi}\int_{\substack{\text{Poincare}\\ \text{sphere}}} (1 - DOP)\, da$

nondepolarizing elements: depolarizing index = 0
ideal depolarizer: depolarizing index = 1

Methods of decomposing depolarizing Mueller matrix

N **non-depolarizing Mueller matrix**

D **"pure depolarizing Mueller matrix"**

a. $M = ND = DN$

b. $M = N + D$

c. $M = \alpha_1 N_1 + \alpha_2 N_2 + \alpha_3 N_3 + \alpha_4 N_4$

How accurate will a polarimeter need to be to measure a specified amount of depolarization?

Which types of samples will be most suitable for initial experimental studies of depolarization?

How is depolarization related to scatering and BRDF?

How will the depolarization vary with the solid angle of light collected about the specular beam?

1991 RESEARCH INITIATION PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT: AFOSR RESEARCH INITIATION PROGRAM– NEURAL NETWORKS FOR GUIDANCE, NAVIGATION, AND CONTROL OF EXOATMOSPHERIC INTERCEPTORS

Prepared by:	Kevin L. Moore, Ph.D.
Academic Rank:	Assistant Professor
Department and University:	College of Engineering Idaho State University
Research Location:	Pocatello, ID 83209
USAF Researcher:	Lt. Kurt Bolin USAF/AFATL/SAI Eglin AFB Ft. Walton Beach, FL 32542

Date:	13 December 1991
Contract No:	F49620-88-C-0053/SB5881-0378

FINAL REPORT: AFOSR RESEARCH INITIATION PROGRAM-NEURAL NETWORKS FOR GUIDANCE, NAVIGATION, AND CONTROL OF EXOATMOSPHERIC INTERCEPTORS

by

Dr. Kevin L. Moore

College of Engineering
Idaho State University
Pocatello, ID 83209

Abstract

Successful deployment of a space-based, exoatmospheric interceptor will require an accurate and reliable guidance, navigation, and control (GNC) system for the interceptor. During the summer of 1990 the proposer began to investigate the application of artificial neural networks to the interceptor control problem under the UES/AFOSR Summer Faculty Research Program. Following the summer research period, AFOSR provided funding to continue this work under the Research Initiation Minigrant Program. The results of the research funded through the Minigrant are described. First, the context of the research grant is given, including a statement of the research objectives and an overview of the results of the research. Next, research results in five specific areas are discussed: (1) an approach to model reference adaptive control for nonlinear systems with characteristics similar to those of the interceptor; (2) development of a control scheme based on reinforcement learning; (3) a method for solving optimal control problems using neural nets; (4) an iterative learning control scheme using artificial neural networks; and (5) an approach to the analysis and design of learning in recursive neural nets using the theory of singular perturbations and time scales. The report ends with a description of future research topics.

Acknowledgements

I wish to thank the Air Force Office of Scientific Research for sponsorship of this research. I would also note the effectiveness of Universal Energy Systems in the administration of the Minigrant program. Their efforts in making the administrative aspects of the program smooth and simple allowed me to concentrate on the technical details of my work.

I would like to acknowledge Lt. Kurt Bolin, the USAF point of contact associated with this project. His enthusiasm and continuing interest in the application of neural nets to the interceptor control problem has been important to the success of the research. I would also acknowledge the role of Dr. D. Subbaram Naidu, of Idaho State University, in a number of fruitful discussions regarding application of the theory of singular perturbation and time scales to neural network problems. Finally, appreciation is extended to Mr. Sudhakar Srinivasan and to my graduate research assistant, Mr. Mark Waddoups, for their assistance with the computer simulations.

Table of Contents

Abstract
Acknowledgements
Table of Contents
I. Introduction
    A. Research Objectives
    B. Research Results
II. Adaptive Control Strategies for GNC
    A. Nonlinear Mappings from Neural Nets
    B. Model Reference Adaptive Control
    C. Adaptive Control of Manipulators
    D. Adaptive Control of GNC Models
III. Reinforcement Learning with Neural Nets
    A. Background
    B. Reinforcement Learning Using Multilayered Nets with Real-Valued Outputs
    C. Example and Comments
IV. Optimal Control Using Neural Nets
    A. Neural Nets for Nonlinear Programming
    B. Optimal Control Solution
    C. Implementation
V. Iterative Learning Control Using Neural Nets
    A. Iterative Learning Control
    B. A Learning Control Scheme
VI. SPATS for Learning in Neural Nets
    A. Singular Perturbations and Time Scales
    B. Hopfield Networks with Learning
    C. An Example of SPATS in a Learning Hopfield Net
VII. Future Research Directions
    A. Learning in Feedforward Neural Nets
    B. Learning in Recursive Neural Nets
    C. Kalman Filtering and Real-Time Adaptive LQR Control

I. Introduction

Successful deployment of a space-based, exoatmospheric guided interceptor is an essential component in the development of strategic space-based weapon systems. Such an interceptor will require an accurate and reliable guidance, navigation, and control (GNC) system. The Guided Interceptor Technology Branch (SAI) of the Air Force Armament Lab (AFATL) at Eglin AFB is concerned with the development of a space-based exoatmospheric guided interceptor. Within SAI the GNC Section is responsible for evaluating advanced GNC technologies for the interceptor. A variety of approaches to interceptor control are currently being studied, including both modern and classical methods. A new technology that has been applied to control problems in other fields is artificial neural networks (ANNs). The GNC Section is interested in determining the potential use of this new technology for the interceptor problem and in comparing its effectiveness to that of the other control schemes under consideration. In this report we consider both basic and applied research related to the control of dynamical systems using ANNs, motivated by the interceptor control problem.

The research reported here is an outgrowth of work completed by the author as a UES/AFOSR Summer Faculty Fellow at AFATL/SAI during the summer of 1990. During the summer research period we began to investigate the application of neural nets to the interceptor control problem. The main results of the summer research period were development of a description of the interceptor control problem and a preliminary analysis of the feasibility of using a neural net to control the interceptor during the terminal phase of flight. This included development of the system configuration, derivation of a math model of the interceptor, and an analysis of current approaches to the GNC problem. Various ways to introduce neural nets into the problem were suggested and it was shown that a feedforward neural network can be used to identify the forward dynamics of the interceptor. A control scheme based on a neural net learning paradigm called reinforcement learning was also proposed. A more complete description of these results can be found in the final report [1].

A. Research Objectives

Based on the results of the summer research period, the following objectives were established to continue the research on ANNs for the interceptor problem. These objectives were taken directly from the Minigrant proposal, hence the future tense in their wording.

1. Model reference adaptive control of nonlinear, multi-input, multi-output systems using neural networks for both identification and control will be studied. This will include additional development of identification of the system's forward dynamics.

2. Our proposed multilayered, real-valued output reinforcement learning scheme will be developed and refined, including theoretical analysis and practical application to single-input, single-output nonlinear systems. This will include work on developing the ability to learn in response to time-varying reinforcement signals.

3. The controllers developed in (1) and (2) will be applied to the interceptor using the GESIM and the resulting performance and robustness properties will be evaluated and compared to that of alternate control schemes.

4. Basic development of neural network implementations of both optimal control and Kalman filtering algorithms will be conducted.

B. Research Results

In the sequel we describe the progress made toward achieving the objectives stated above. The description is divided into the following five parts:

1. Adaptive Control Strategies for GNC: An approach to adaptive control for nonlinear systems with models similar to those of the interceptor is presented. The method is a derivative of the computed-torque method found in robotic manipulator control.

2. Reinforcement Learning with Neural Nets: The reinforcement learning scheme proposed during the summer of 1991 has been developed. It is demonstrated for a single-input, single-output example.

3. Optimal Control Using Neural Nets: A neural network approach to solving linear quadratic regulator problems is given.

4. Iterative Learning Control Using Neural Nets: A method for using neural networks to devise a dynamical iterative learning controller is presented.

5. SPATS for Learning in Neural Nets: We discuss some preliminary results which interpret recursive Hopfield neural nets from the perspective of singular perturbations and time scales.

Items (1) through (3) correspond to items (1),(2), and (4) of the objectives, respectively. For these items we have been generally successful in meeting the stated objectives, although simulations of the techniques we have developed are not complete for items (1) and (3). However, we were not able to complete item (3) of the objectives: the application of our control schemes to the interceptor using GESIM (the software at Eglin AFB that is used to simulate the interceptor). The methods were not sufficiently developed soon enough to

incorporate them into the software. This will require additional research and development effort.

Also note that items (4) and (5) of the results were not specifically tasked in the stated objectives. However, research as a creative activity is a fluid affair, acting to fill the space around it and often following many paths. This is distinct from development, which often has product-oriented goals. In a broad sense we view our Minigrant as supporting the general problem of control of nonlinear systems using neural nets (which encompasses the interceptor control problem). Items (4) and (5) deal with aspects of this general problem and were motivated by questions raised by our Minigrant activity. Because they were completed during the time of the Minigrant funding, we see these items as indirectly part of the Minigrant and feel they should be included. Although it is unlikely that item (4) will ever be applied to the interceptor problem, it will be useful in some types of robotics problems, and as such, it may be of interest in some AFOSR programs. However, in the long term, item (5) may in fact prove useful to the interceptor problem.

In the remainder of the report we separately describe each of the results stated above. For the most part, each section is self-contained and generic. That is, if a given approach applies to nonlinear control in general, then it is presented in that way. Much of this work has been reported as it was developed. The work on adaptive strategies for GNC has not yet been published. The work on reinforcement learning was presented in [2]. The work on optimal control has been reported at a number of meetings, notable [3] and [4]. The iterative learning control technique has been submitted to a conference [5]. The work on the application of singular perturbations and time scales is presented in [6]. Finally, a result on system identification that we refer to in the final section is given in [7]. This work was initiated by a student prior to the Minigrant funding and hence it is not included in this final report.

II. Adaptive Control Strategies for GNC

In this section we present some preliminary ideas about using artificial neural nets (ANNs) as elements in adaptive control systems for guidance, navigation and control (GNC) applications. We first describe how the nonlinear mapping property of an ANN can be used to produce dynamic behavior. We then consider model reference adaptive control in general. Next we describe the computed-torque method used for adaptive control of robotic manipulators. Finally, we show how a similar scheme can be developed for typical GNC models.

A. Nonlinear Mappings from Neural Nets

As noted in [1], one of the most appropriate uses of a neural net for control applications is as a nonlinear mapping. This is motivated by Kolmogorov's result from the late 1950's which states that neural net structures can be used to implement arbitrary functions (although there is no indication how these functions can be learned, i.e., it is an existence result, and of course, certain restrictions apply). The reason this is a good approach to neural nets for control is that most control systems are designed for dynamical systems which can be described by equations such as

$$y_{k+1} = f(y_k, y_{k-1}, y_{k-2}, u_k, u_{k-1})$$

In this discrete-time system we see that the next output is a function of the current input, the current output, the most recent past input, and the two most recent past outputs. Thus we could imagine the configuration shown in Figure 1, where the operator z^{-1} denotes a unit time delay and the neural net implements the nonlinear mapping f [8]. This leads naturally to several approaches to control. Below we describe one of those approaches: model reference adaptive control.

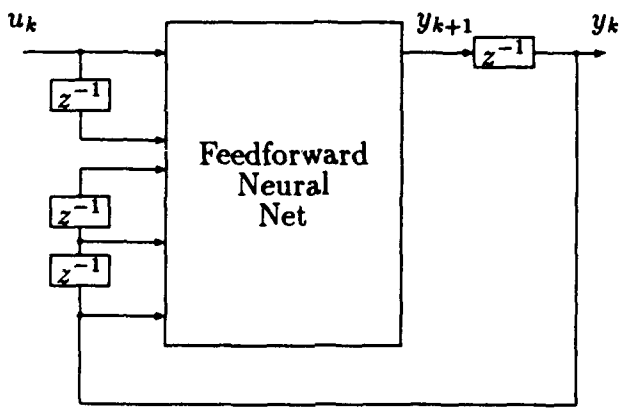


Figure 1: Neural net implementation of discrete-time system.

B. Model Reference Adaptive Control

Recently Narendra and Parthasarathy have presented an approach for the use of neural networks within the framework of conventional model reference adaptive control (MRAC) [8]. Consider, for instance, a nonlinear system (example taken from Narendra's paper) of the form

$$y_{k+1} = f(y_k, y_{k-1}) + u_k$$

where

$$f(x, y) = \frac{xy(x + 2.5)}{1 + x^2 + y^2}$$

The goal of the control is to match the dynamics of a reference model given by

$$y_{k+1} = .6y_k + .2y_{k-1} + r_k$$

where r_k is the reference signal. Suppose the control law is

$$u_k = -N(y_k, y_{k-1}) + .6y_k + .2y_{k-1} + r_k$$

where $N(x, y)$ is a feedforward neural network which is identified using backpropagation training. It can be seen that if the net successfully learns the system dynamics then the output behavior will approach the desired behavior because the closed-loop system is now given by

$$y_{k+1} = f(y_k, y_{k-1}) - N(y_k, y_{k-1}) + .6y_k + .2y_{k-1} + r(k).$$

Narendra has considered various examples of this type, showing how neural nets can be used to identify nonlinear dynamics, with the results of the identification then used as part of a control law. Although most of this work is demonstrated for single-input, single-output systems, it gives a very good foundation for the use of neural nets in control. In particular, notice that this approach allows us to account for the temporal characteristics of the system using a feedforward neural network, which is typically only used for identifying static patterns.

C. Adaptive Control of Manipulators

We are interested in how a model reference adaptive control scheme using neural networks can be applied to GNC problems. One feature of Narendra's approach is that it uses knowledge of the system's structure to configure the neural net control system. In order to consider model reference control for the interceptor, we must first consider the structure of the interceptor's equations of motion. When we do this (see below), we notice that the dynamical equations of the interceptor are similar in structure to those that arise in robotic manipulator control. Thus it is reasonable to consider the adaptive control schemes that have been applied in robotics.

The equations of motion for an n-jointed manipulator can be written as

$$\begin{aligned} \dot{\theta}_1 &= \theta_2 \\ \dot{\theta}_2 &= T(\theta_1, \theta_2) + B(\theta_1, \theta_2)u \end{aligned}$$

where $\theta_1(t)$ is the vector of joint angles; $\theta_2(t)$ is the vector of joint angular velocities; $T(\theta_1, \theta_2)$ and $B(\theta_1, \theta_2)$ are matrices whose elements are nonlinear functions of θ_1 and θ_2; and u is the torque vector applied to the joint actuators. It is assumed that B has full rank. By making the substitution $\theta_2 = \dot{\theta}_1$ we obtain

$$\ddot{\theta}_1 = T(\theta_1, \dot{\theta}_1) + B(\theta_1, \dot{\theta}_1)u$$

Often this equation is further reduced to the form

$$M(\theta_1)\ddot{\theta}_1 + V(\theta_1, \dot{\theta}_1) + G(\theta_1) = u$$

where $M(\theta_1)$ is identified as the inertial mass matrix; $V(\theta_1, \dot{\theta}_1)$ represents centrifugal and coriolis forces; and $G(\theta_1)$ represents the effect of gravitational forces [9].

A popular approach for adaptive control of an n-jointed manipulator with the model described above is the so-called computed-torque method. This scheme assumes a control law given by

$$u = \hat{M}(\theta_1)(\ddot{\theta}_d + K_v\dot{e} + K_p e) + \hat{V}(\theta_1, \dot{\theta}_1) + \hat{G}(\theta_1)$$

In this equation θ_d is the desired joint position trajectory, $e = \theta_d - \theta_1$ is the measured joint error, and K_v and K_p are diagonal gain matrices. In this case, if an appropriate identification scheme is developed so that $\hat{M} \to M, \hat{V} \to V$, and $\hat{G} \to G$, then the closed loop system is described by

$$\ddot{e} + K_v\dot{e} + K_p e = 0.$$

Thus, it is possible to drive the joint error to zero if the identification proceeds properly. This scheme has been shown to be successful both theoretically and experimentally.

One idea for using ANNs to control manipulators is to use neural nets to identify the matrices M, V, and G. The nets would then be used in the computed-torque law to compute the required input torque needed to achieve a desired motion. Below we suggest a similar approach for the interceptor.

D. Adaptive Control of GNC Models

The equations of motion that arise in many GNC problems are similar to those describing an n-jointed manipulator. However, due to the presence of both inertial reference frames and body-fixed frames, there are some differences. Specifically, usually $\theta_2 \neq \dot{\theta}_1$. For instance, in [1] it is noted that the interceptor model has the general form

$$\begin{aligned} \dot{\theta}_1 &= T_1(\theta_1)\theta_2 \\ \dot{\theta}_2 &= T_2(\theta_2)\theta_2 + Bu \end{aligned}$$

For systems modelled in this way we can still proceed, if it is possible to assume that (i) $T_1^{-1}(\theta_1)$ exists and (ii) the partial derivative of T_1 with respect to time is well-defined. In this case we proceed as follows. Given the system above, then $\theta_2 = T_1^{-1}(\theta_1)\dot{\theta}_1$ and

$$\begin{aligned} \ddot{\theta}_1 &= \frac{\partial T_1}{\partial t}(\theta_1)\theta_2 + T_1(\theta_1)\dot{\theta}_2 \\ &= \frac{\partial T_1}{\partial t}(\theta_1)\theta_2 + T_1(\theta_1)T_2(\theta_2)\theta_2 + T_1(\theta_1)Bu \\ &= \left[\frac{\partial T_1}{\partial t}(\theta_1) + T_1(\theta_1)T_2(T_1^{-1}(\theta_1)\dot{\theta}_1)\right] T_1^{-1}(\theta_1)\dot{\theta}_1 + T_1(\theta_1)Bu \end{aligned}$$

We can then write

$$M(\theta_1)\ddot{\theta}_1 + V(\theta_1, \dot{\theta}_1) = u$$

where we can identify

$$\begin{aligned} M(\theta_1) &= [T_1(\theta_1)B]^{-1} \\ V(\theta_1, \dot{\theta}_1) &= [T_1(\theta_1)B]^{-1}\left[\frac{\partial T_1}{\partial t}(\theta_1) + T_1(\theta_1)T_2(T_1^{-1}(\theta_1)\dot{\theta}_1)\right]T_1^{-1}(\theta_1)\dot{\theta}_1 \end{aligned}$$

This is now in essentially the same form as the system equations of the robotic manipulator. Thus we may consider the following variation of the computed torque method. Let

$$u = \hat{M}(\theta_1)(\ddot{\theta}_d + K_v\dot{e} + K_p e) + \hat{V}(\theta_1, \dot{\theta}_1)\dot{\theta}_1.$$

Again we have defined θ_d as the desired position trajectory, $e = \theta_d - \theta_1$ as the measured error, and K_v and K_p as diagonal gain matrices. In this case, if an appropriate identification scheme is developed so that $\hat{M} \to M$ and $\hat{V} \to V$ then the closed loop system will described by

$$\ddot{e} + K_v\dot{e} + K_p e = 0.$$

In order for such a scheme to work a number of issues must be resolved. Most importantly, we must be sure of an identification scheme that will converge. Based on our preliminary results with a six degree-of-freedom model [1], we believe that a neural network can do this job. Other issues involve the existence of the inverses used in the development above. Current work is concentrating on establishing that the assumptions are in fact satisfied for realistic problems, such as the interceptor. We are also conducting simulations of the method for a simple two-dimensional system given by

$$\begin{aligned} \dot{\theta}_1 &= (\cos\theta_1 + \sin\theta_2)\theta_2 \\ \dot{\theta}_2 &= \theta_2^2 + u. \end{aligned}$$

III. Reinforcement Learning with Neural Nets

Reinforcement learning in neural nets is an approach to the problem of credit assignment during learning. As opposed to gradient descent techniques such as backpropagation, a reinforcement learning scheme uses a single reinforcement signal from the environment to adjust the network weights. In this section we describe reinforcement learning and propose a multilayer neural network with real-valued outputs which learns using a combination of reinforcement learning and backpropagation. This method combines several ideas from the literature. We illustrate the use of the method with an example of the control of a nonlinear system. This develops and extends our previous work in this area [1].

A. Background

A key issue in the development of a neural network application is the type of learning algorithm that is used. Most algorithms currently being studied, such as backpropagation, are gradient descents in the weight space. This is one approach to what has been called the "credit assignment" problem. That is, how to reinforce weights that contribute to desirable behavior of the system. Another approach to this credit assignment problem is to use a single reinforcement signal from the environment (system) to decide how to adjust the weights. This is decidedly different from backpropagation, which essentially attempts to form an error signal at the output of each neuron in the network. In reinforcement learning we have only one error signal, which is used by all the neurons.

To describe reinforcement learning consider Figure 2. Here we suppose the neural net

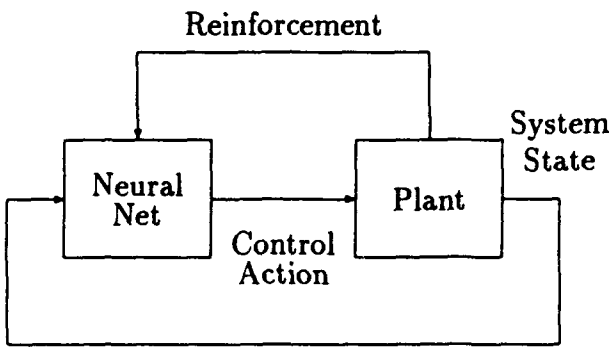


Figure 2: Reinforcement learning.

generates control actions, based on measurements of the current system state. It then receives a reinforcement signal from the environment which evaluates the effect of the action. The weights are then adjusted based upon the reinforcement. In the learning algorithm, the only information given to individual neurons, regardless of their location in the net, is the reinforcement signal. There is no error that is backpropagated through the net. In a typical algorithm, only those weights which contributed to the action of the net will be adjusted. Specifically, there is an eligibility function associated with each weight which indicates how recently the weight has been changed or how recently the neurons associated with a weight have fired. Only weights that are "eligible" are updated during learning. Learning initially proceeds randomly, but as the net learns which actions lead to desirable reinforcement in response to specific system states, the weights begin to adjust in such a way that a bias is built up toward these actions. Thus this type of system is in fact a stochastic automata.

Past work in reinforcement learning followed work in stochastic automata theory [10] and psychology [11]. More recently, a research group based around Barto and Sutton has been very active. Initial results on using reinforcement learning control for the pole balancing problem can be found in Barto, Sutton, and Anderson [12]. A key feature of this work was that it used an input decoder to partition the state space into "boxes". They used a network made up of what they called "neuron-like" elements which were basically perceptrons with stochastic outputs. The network applied control forces to a pole-cart system in an attempt to balance the pole. Learning (adjustment of the weights, based on a reinforcement signal from the system) took place only when the pole fell over. The system was able to learn which actions would keep the pole balanced. They also introduced the idea of trying to predict the value of the reinforcement signal. This allowed faster learning. One limitation in the work of Barto, Sutton, and Anderson was that it allowed only binary outputs. Franklin [13] was able to extend their ideas to allow real-valued outputs and has applied this to the control of a simulated robot arm. Anderson [14] has also considered the use of multilayer networks with reinforcement learning. Again, this work uses binary outputs, but the input decoder is replaced by a neural net which learns to distinguish the system state without quantizing it into boxes. Also, the learning algorithms are based on the theory of temporal differences which has been introduced by Sutton [15] and include a combination of reinforcement learning theory and gradient descent rules.

B. Reinforcement Learning Using Multilayered Nets with Real-Valued Outputs

Figure 3 shows a proposed configuration we have developed for the control of a nonlinear system. This scheme is a combination of Franklin's ideas and Anderson's ideas. The system uses multilayer neural nets with real-valued outputs. Specifically, there are two parts of the control system: an action part and a predictor. The control output (u) is a normal random variable with mean equal to the output of the action net (μ) and with standard deviation given by the predicted reinforcement (σ). The idea is that if the predictor expects the proposed action μ to result in a large (bad) reinforcement signal then the actual control action is computed as a random variable centered at μ, but with a large variance. Thus the system is less likely to take an action close to μ. On the other hand, if σ is small this means the predictor expects a more favorable reinforcement and thus allows the control action to be closer to μ.

Both the action net and the predictor net are standard feedforward nets with outputs defined by

$$
\begin{aligned}
z &= \sum_i c(i)u(i) + \sum_j b(j)v(j) \\
v(j) &= f(\sum_k a(j,k)w(k))
\end{aligned}
$$

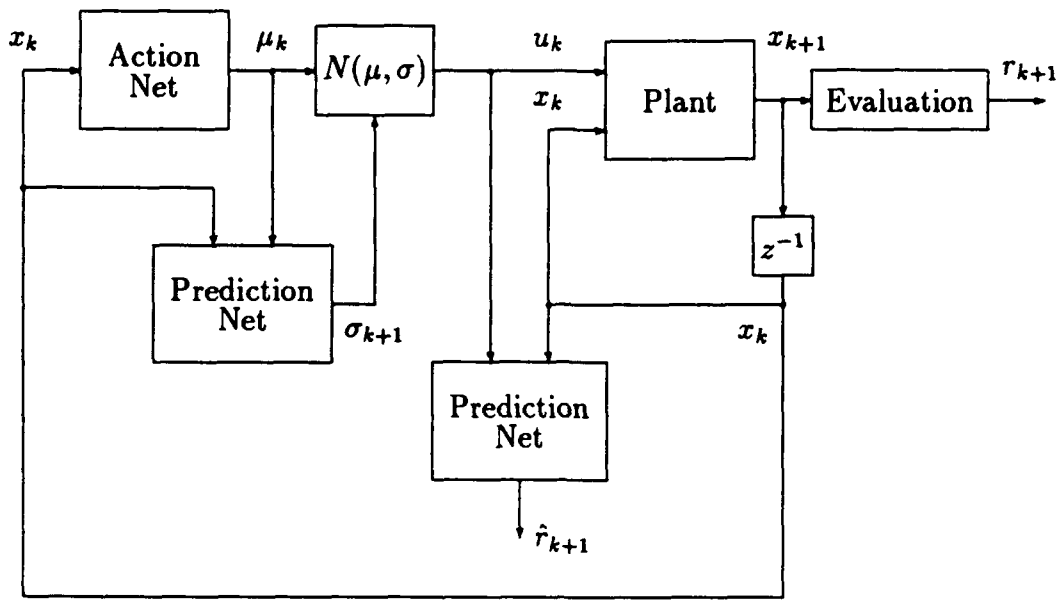


Figure 3: Proposed control scheme.

where z is the net's output, the $v(j)$'s are the outputs of the hidden layer, the $w(i)$'s are the inputs to the net, and $f(\cdot)$ is a sigmoidal nonlinearity. Note that in our scheme, the action net uses only the system states (x_k) as inputs. The predictor net receives as inputs the system states and either the control action (u) or the expected control action (μ), depending on whether the net is being used in a learning mode or in an operational mode, respectively.

The weights of the predictor net are continuously adjusted using the error between the output $\hat{r}_{k+1}$ and the actual reinforcement r_{k+1}. Thus the predictor learns the reinforcement resulting from a given action (u_k) within a given context (x_k). Training of this net is done using standard backpropagation with an error signal $e = r - \hat{r}$. This gives

$$\begin{aligned} b(i)_{new} &= b(i)_{old} + \eta_b e v(i) \\ c(i)_{new} &= c(i)_{old} + \eta_c e w(i) \\ a(i,j)_{new} &= a(i,j)_{old} + \eta_a e[b(i)v(i)(1 - v(i))]w(j) \end{aligned}$$

Once the weights of the predictor are adjusted on a given trial, it is then used in an operational mode to compute the variance σ_{k+1} used for computing the next control action. Note that after learning we have $\sigma_{k+1} = \hat{r}_{k+1} = r_{k+1}$.

Learning in the action layer is a combination of conventional backprop and reinforcement learning. The same type of backpropagation equations given above for the predictor are used,

except that the error signal is modified to be

$$e = R_p \cdot (u - \mu)$$

where

$$R_p = \begin{cases} 1 & \ldots \text{ if } \sigma_{k+1} > r_{k+1} \\ 0 & \ldots \text{ otherwise} \end{cases}$$

R_p is a flag which indicates a desirable system response to the most recent action. Specifically, this signal provides a mechanism for positive reinforcement. The idea behind this algorithm is the following. Suppose that a given control action u results in an actual reinforcement that is smaller than the predicted reinforcement. Then we should adjust the weights so that for the same set of inputs the mean signal (μ) is moved closer to the actual control action u that gave a better reinforcement. However, if the reinforcement is larger than expected, then we make no changes, but simply allow the stochastic search for a better action to proceed on the next trial.

C. Example and Comments

This scheme was applied to the following example plant from Narendra's paper [8].

$$y_{k+1} = \frac{y_k y_{k-1}(y_k + 2.5)}{1 + y_k^2 + y_{k-1}^2} + u_k.$$

In our example, the goal was to train the system output to follow a reference signal. In order to do this the configuration of Figure 3 was modified to allow input of the desired reference signal to each net. The reinforcement signal for the example was taken to be the error between the system output and the reference signal. The predictor provided an estimate of this error. Training of the action net occurred whenever the actual error was smaller than the predicted error. In general the scheme gave good results. Although the net architecture is somewhat different, the predictor could identify the system dynamics with the same results as the example in Narendra's paper. We also found the control action to be successful. Figure 4 shows a sample result after about 20,000 time steps for the case of a square wave reference signal. In this trial both the predictor net and the action net had five hidden layer nodes and all the learning gains were set to be 0.01.

We feel there are several aspects of reinforcement learning schemes which make them potentially useful for the control of the nonlinear systems. In particular, because the training is stochastic in nature, the neural net will have learned to respond to a variety of input conditions. Thus we expect it to be fairly robust with respect to parameter variations or unmodelled dynamics. We are encouraged by our preliminary results, but more work remains to be done. We are particularly interested in how these ideas could be applied to response shaping and we are currently working to develop a system which learns in response to a

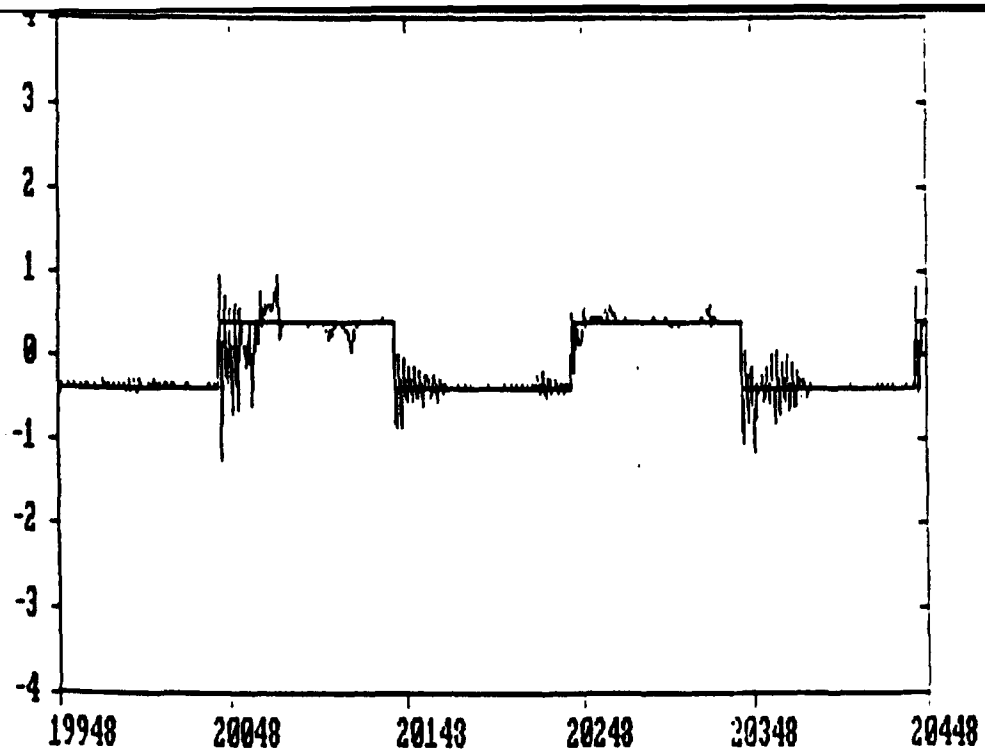


Figure 4: Tracking response.

time-varying reinforcement signal. Such a signal could be bounded above and gradually grow to zero to reflect overshoot and settling time requirements, respectively. We are also in the process of performing a detailed analysis of the behavior of the proposed control scheme, including additional simulations of a two-input, two-output system. Our goal is to realize a successful application of the scheme to the full six degree-of-freedom model of the interceptor.

IV. Optimal Control Using Neural Nets

In this section we describe the use of artificial neural networks for solving optimal control problems for discrete-time linear systems with quadratic cost functions. The result is obtained by formulating the optimal control problem as a quadratic programming problem with inequality constraints and then applying a result by Kennedy and Chua [16]. We first describe the result that allows us to use a neural net to solve optimization problems. Then we show how to formulate the linear quadratic regulator (LQR) problem as a nonlinear programming problem. It is then possible to directly apply Kennedy and Chua's result to find the optimal control solution using a neural net.

A. Neural Nets for Nonlinear Programming

Consider the general nonlinear programming problem with inequality constraints:

$$\min_{v \in R^n} \phi(v)$$

subject to

$$f_i(v) \geq 0$$

for $i = 1, \ldots, q$. All the standard assumptions about constraint qualification, continuity, existence of first and second partials, etc., are assumed to apply. It has been shown that canonical nonlinear programming circuit of Kennedy and Chua solves this nonlinear programming problem [16]. The circuit, shown in Figure 5, is in the form of a Hopfield neural network and is described by the following equations.

$$C_i \frac{dv_i}{dt} = -\frac{\partial \phi}{\partial v_i} - \sum_{j=1}^{p} g_j(f_j(v)) \frac{\partial f_j}{\partial v_i},$$

for $i = 1, \ldots, q$. Here the $g_j(\cdot)$'s are nonlinear resistors which are required to be passive monotone nondecreasing functions.

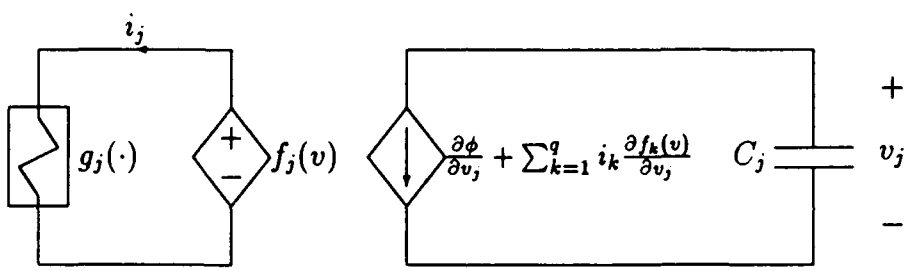


Figure 5: Neural network for optimization.

B. Optimal Control Solution

Optimal control problems for linear systems with quadratic cost functions have been described in a variety of references. For a typical development see Kailath [17]. In this section we restrict our analysis to the discrete-time, deterministic problem. Specifically, we consider the following finite-horizon problem.

$$\min_{u_k} J = \frac{1}{2} x_N^T S x_N + \frac{1}{2} \sum_{k=0}^{N-1} \{x_k^T Q x_k + u_k^T R u_k\}$$

subject to the plant dynamics

$$x_{k+1} = A x_k + B u_k$$

for $k = 0, \ldots, N-1$, with x_0 given. We assume $x \in R^n$ and $u \in R^p$. Also assume the pair (A, B) is controllable.

Now, define a "supervector" $v \in R^{(N+1)n+Np}$ by

$$v = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_N \\ u_0 \\ u_1 \\ \vdots \\ u_{N-1} \end{pmatrix}$$

Then notice that the optimal control problem can be written more compactly as

$$\min_v J$$

subject to

$$\Gamma v = 0.$$

where

$$\Gamma = \begin{bmatrix} -A & I & 0 & 0 & \cdots & 0 & 0 & -B & 0 & \cdots & 0 & 0 \\ 0 & -A & I & 0 & \cdots & 0 & 0 & 0 & -B & \cdots & 0 & 0 \\ 0 & 0 & -A & I & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & I & 0 & 0 & 0 & \cdots & -B & 0 \\ 0 & 0 & 0 & 0 & \cdots & -A & I & 0 & 0 & \cdots & 0 & -B \end{bmatrix}$$

Note that this optimization problem involves only equality constraints. In order to apply the result of Kennedy and Chua it is necessary to transform the Nn equality constraints into $Nn+1$ inequality constraints by applying the following result from Taha [18].

Theorem Given m equality constraints,

$$\sum_{j=1}^{n} a_{ij} x_j = b_i$$

for $i = 1, \ldots, m$, these are equivalent to the following m inequality constraints

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i$$

for $i = 1, \ldots, m$, along with the additional inequality constraint

$$\sum_{j=1}^{n} (\sum_{i=1}^{m} a_{ij}) x_j \geq \sum_{i=1}^{m} b_i.$$

That is, we add a constraint involving a column sum of the elements of the A matrix.

Using this fact we can now rewrite the LQR problem in the following form

$$\min_{v} J$$

subject to

$$f_i(v) = \sum_{j=1}^{(N+1)n+Np} \bar{a}_{ij} v_j \geq 0$$

for $i = 1, \ldots, Nn+1$, where the $\bar{a}_{ij}$'s are elements of the matrix $\bar{A}$, defined by

$$\bar{A} = \begin{bmatrix} -A & I & 0 & \cdots & 0 & 0 & -B & \cdots & 0 & 0 \\ 0 & -A & I & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -A & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & I & 0 & 0 & \cdots & -B & 0 \\ 0 & 0 & 0 & \cdots & -A & I & 0 & \cdots & 0 & -B \\ \alpha^T & (\alpha^T - \underline{1}) & (\alpha^T - \underline{1}) & \cdots & (\alpha^T - \underline{1}) & \underline{1} & \beta^T & \cdots & \beta^T & \beta^T \end{bmatrix}$$

In this matrix α, $\underline{1}$, and β are defined as the following.

$$\begin{aligned} \alpha &= (\sum_{i=1}^{n} a_{i1}, \sum_{i=1}^{n} a_{i2}, \ldots, \sum_{i=1}^{n} a_{in})^T \\ \underline{1} &= (1, 1, \ldots, 1)^T \\ \beta &= (\sum_{i=1,n} b_{i1}, \sum_{i=1}^{n} b_{i2}, \ldots, \sum_{i=1}^{n} b_{ip})^T \end{aligned}$$

Using this formulation of the problem we now have a direct correspondence between the optimal control problem and the nonlinear programming problem, which can be solved with the neural net given above.

C. Implementation

Given the problem as formulated above it is necessary to map the constraints and all the appropriate derivatives into the neural network circuit. This involves computing

$$\frac{\partial J}{\partial v_i}$$

and

$$\frac{\partial f_i}{\partial v_j}.$$

To compute the partials of f_i with respect to the variable v_j, the easiest thing to do is first form the matrix $\bar{A}$. Then the derivative of the inequality f_i with respect to the variable

v_j is simply the $(ij)^{th}$ element of $\bar{A}$ (that is, $\bar{a}_{ij}$). However, the matrix $\bar{A}$ has some definite structure that can add insight into the neural net implementation of the equations. A topic for future study is the nature of this structure and its implications.

To compute the partials of the cost functional J with respect to the variable v_j, first suppose that the weight matrices Q, S, and R are defined as

$$Q = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & \ddots & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix}$$

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix}$$

and

$$R = \begin{bmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{bmatrix}.$$

Then it is straightforward to show the following.

$$\frac{\partial J}{\partial v_i} = \begin{cases} \sum_{j=1}^{n} q_{i-ln,j} v_{ln+j} & \ldots \quad i = ln+1, \ldots, (l+1)n \\ & \qquad\quad l = 0, \ldots, N-1 \\ \sum_{j=1}^{n} s_{i-Nn,j} v_{Nn+j} & \ldots \quad i = Nn+1, \ldots, (N+1)n \\ \sum_{j=1}^{n} r_{i-(N+1)n,j} v_{(N+1)n+j} & \ldots \quad i = (N+1)n+1, \ldots, (N+1)n+p \end{cases}$$

where q_{ij}, s_{ij}, and r_{ij} are the $(ij)^{th}$ elements of the matrices Q, S, and R, respectively. As in the case of the partials of J with respect to the v_j's, there is some obvious structure inherent in these equations which we plan to study in the future.

We are currently working to develop this approach to optimal control. One focus has been to develop some numerical examples of the technique using P-SPICE to simulate the net. To date this has not been successful, primarily because of hardware and software limitations. However, we do expect to obtain successful simulations in the near future. This will allow us to compare the method to conventional solutions of optimal control problems. Other directions for extending these ideas are detailed in the final section below.

V. Iterative Learning Control Using Neural Nets

In this section we describe some results which apply neural nets to the problem of iterative learning control. Although this was not specifically tasked as part of the objectives of the

Minigrant, we report these results because they were obtained during the period of the Minigrant research. Also, we were led to reconsider the iterative learning control problem as a result of our work on the reinforcement learning scheme, which is a stochastic type of iterative learning controller. Further, these results represent basic research in the area of neural network control of nonlinear dynamical systems, an area which encompasses the interceptor control problem.

A. Iterative Learning Control

A new approach to the problem of designing for specific transient response performance is the approach of iterative learning control. Originally suggested and developed by Arimoto and his co-workers [19], learning control is an iterative approach to generating the optimal system input so that the system output is as close as possible to the desired output. The method can be used to improve the transient behavior of systems that are repetitive in nature and operate over a fixed time interval. An example of such a system is an antenna servomechanism in a radar which performs a given scan pattern over and over. The trajectory which the antenna follows is completed in a finite time and then it is repeated. Learning control is also well-suited to problems in which a system must be able to follow different types of inputs. For example, in a flexible manufacturing system a robot may be placed on an assembly line and programmed to paint a straight line on a car door as it passes the robot's station. Later, as production needs change, the manufacturing facility may be reconfigured and perhaps the robot is needed to execute a circular trajectory to paint whitewall tires. Thus the robot is required to execute a given trajectory many times for a specific task, but may be configured to perform several different tasks, depending upon production needs.

To understand the approach of learning control consider Figure 6. Each time the system

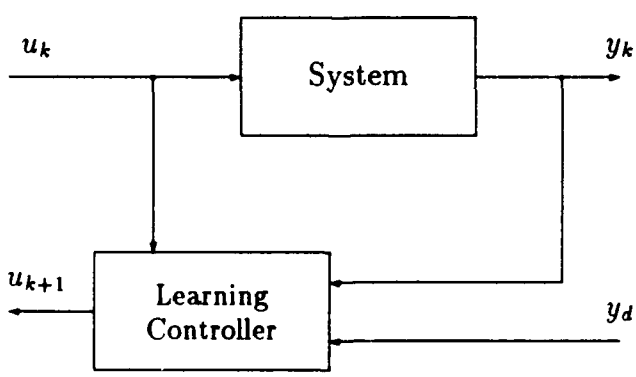


Figure 6: Learning control configuration.

operates the input to the system, u_k, is stored, along with the resulting system output, y_k. That is, we store the complete trajectory over the entire interval of interest. The learning controller then evaluates the performance error as compared to some desired signal y_d and computes a new input, u_{k+1}, which is stored for use the next time the system operates. The new input is chosen in such a way as to guarantee that the performance error will be reduced on the next trial. The task in learning control is to specify the algorithm for generating the next input, given the current input and output, so that convergence to the desired output (in the sense of some norm) is attained. Ideally, the convergence property of the learning control algorithm would require a minimal knowledge of the system parameters and would be independent of the desired response y_d. Thus it would be possible to derive the best control signal for any of several desired responses, in the presence of modelling uncertainty, simply by using input-output data obtained from actual operation. Note this key difference between learning control and optimal control design techniques. In conventional optimal control methodologies the optimal input is computed *a priori* using a model of the system. Also note that learning control differs from conventional adaptive control in that the learning algorithm is executed off-line, at the end of each trial. Most adaptive control schemes are on-line algorithms and involve changes of the controllers parameters. Because of its iterative nature, learning control can be described as training. Given a desired response, we train the system to follow it with as little transient error as possible.

B. A Learning Control Scheme

In previous work [20] we have noted that the real usefulness of iterative learning control may be for problems in which we wish to control the transient response behavior of nonlinear or time-varying systems. In this case it makes sense to consider learning controllers which also have a nonlinear or time-varying structure. A non-trivial question, however, is what type of nonlinear system should be considered. The class of all nonlinear systems is very large and it is not clear what structure would work best for learning control applications. One answer to this question is to consider learning controllers based on artificial neural networks. ANNs, with their nonlinear structure and their ability to learn internal representations and recall associations, are good candidates for nonlinear learning control.

Previously we have developed a method for learning control which required the same number of neurons in both the input layer and the output layer as the number of time steps in the interval of interest [20]. Here we propose to use the approach of Narendra and Parthasarathy to develop a nonlinear learning controller which takes into account the dynamic relationship between the input/output data. The advantage of this method over our earlier technique is that the number of inputs to the neural net of the learning controller need be no greater than twice the order of the plant. This is true no matter how large we

make N, the length of the learning control trials. In contrast, our previous method would require N inputs and N outputs. This will have convergence problems as N increases.

The proposed learning control scheme is shown in Figure 7. We call this a dynamical learning control scheme because the final learning controller actually acts like a dynamical system, using past outputs as inputs.

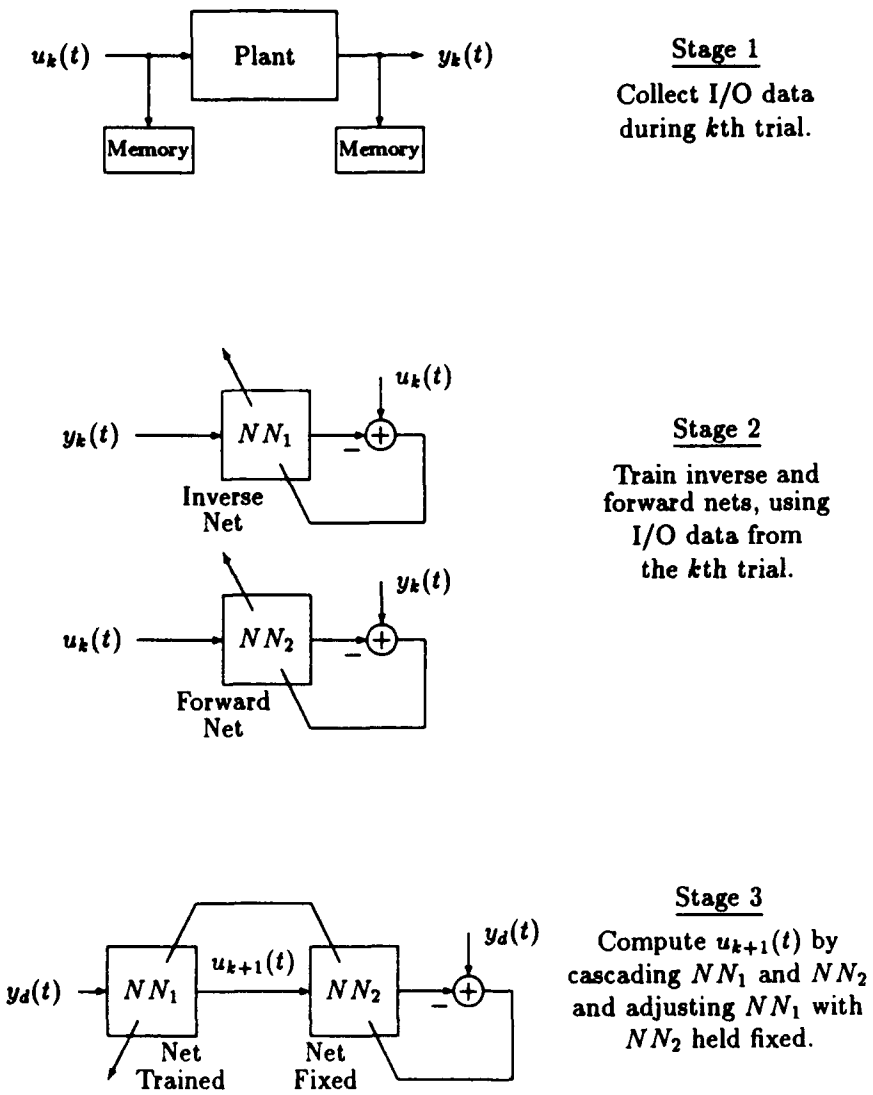


Figure 7: Dynamical learning controller.

Notice that in this system there are two neural networks. One, NN_2, is used to identify the dynamics of the plant. The other, NN_1, is used to identify the inverse dynamics of the plant. It is assumed that the order of the plant is known and that each neural net block shown in the figure is actually configured as in Figure 1. The reason for using two nets rather

than just one net is to avoid the problem of projecting through a gradient of the plant to get the correct error for learning the inverse dynamics.

The following procedure is used to compute a new input to the system at the end of each trial.

1. Set the initial input $u_0(t) = y_d(t)$ and set $k = 0$.

2. Stage 1: Run the plant with u_k. Collect and store in memory $u_k(t)$ and $y_k(t)$.

3. Stage 2: Parallel learning. Train the inverse and forward nets using the input/output data from the kth trial. Training is on a point-by-point basis, using backpropagation.

4. Stage 3: Series learning. Cascade the inverse and forward nets. Apply the desired output $y_d(t)$ as the input to the inverse net. Train to force the output of the forward net to match the desired output. Use backpropagation, but hold the forward net fixed. Only the inverse net is adjusted.

5. Series learning and parallel learning are alternated until convergence of error signals.

6. Recover the input for the next trial, u_{k+1} from the output of the inverse net after learning is complete.

7. Set $k = k + 1$ and GOTO Step 2 (Stage 1).

A key aspect of this algorithm is the alternation between Stage 2 and Stage 3 during learning. This ensures that we have adjusted the net at each trial to invert our best approximation to the plant (in the direction of the desired output) while preserving the integrity of the inverse model for the actual input/output data.

To illustrate the technique, we simulated the scheme for the plant mentioned above, from Narendra and Parthadarathy's paper [8].

$$y_{k+1} = \frac{y_k y_{k-1}(y_k + 2.5)}{1 + y_k^2 + y_{k-1}^2} + u_k.$$

The performance of the learning control scheme is shown in Figure 8 (the desired signal is a triangular waveform). As can be seen, within about ten trials the actual performance matches the desired performance almost exactly.

VI. SPATS for Learning in Neural Nets

In this section we describe an approach we have developed for the analysis and design of learning in neural nets and in systems containing neural nets as subsystems. Throughout, we restrict our attention to recursive ANNs, such as the Hopfield net. However, much of

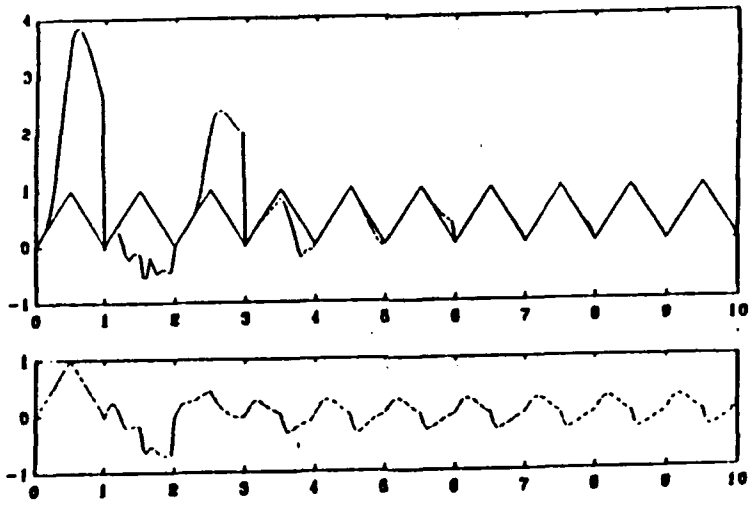


Figure 8: Dynamical learning controller performance: Top graph – desired and actual response. Bottom graph – control input signal.

the discussion also applies to feedforward nets operated in feedback configuration. The approach introduced in this section is to apply well-developed techniques from the theory of singular perturbations and time scales (SPATS) to standard neural network architectures. The basic idea is to cast a neural network system into the standard framework of SPATS by identifying the "learning" and "computing" processes in the neural net as "slow" and "fast" phenomena, respectively. As in the previous section, this topic was not explicitly tasked in the objectives of the Minigrant. Again, however, these ideas were developed during the period of the Minigrant and we feel they are relevant to the interceptor control problem from the context of their relevance to the general problem of control of nonlinear systems using neural networks.

A. Singular Perturbations and Time Scales

Consider a (usually nonlinear) dynamical system described by

$$\begin{aligned} \frac{dx_1}{dt} &= f(x_1, x_2) \\ \epsilon\frac{dx_2}{dt} &= g(x_1, x_2, u) \end{aligned}$$

with external input u, boundary conditions $x_1(0)$ and $x_2(0)$, and ϵ a small positive parameter. Systems which can be described in this way are said to be singularly perturbed as a result of the parameter ϵ. Typically we identify ϵ as the time constant of the "fast" state

x_2 and we identify x_1 as the "slow" state of the system. This description of the system into fast and slow modes results in multiple time scales for use in describing the system behavior. Other phenomena associated with such systems include order reduction, loss of some initial conditions, and boundary layer formation (region of rapid transition). The theory of singular perturbations and time scales (SPATS) is well-developed [21]. There are three aspects of the theory that we will mention: degeneration, asymptotic expansion, and time-scale decomposition.

(i) *Degeneration*: By setting the small parameter $\epsilon = 0$ we obtain the degenerate (low-order) problem

$$\begin{aligned} \frac{dx_1^{(0)}}{dt} &= f(x_1^{(0)}, x_2^{(0)}) \\ 0 &= g(x_1^{(0)}, x_2^{(0)}, u) \end{aligned}$$

where $x_1^{(0)}(t)$ and $x_2^{(0)}(t)$ represent the degenerate solutions of the slow and fast modes of the system, respectively. In general, because the order has been reduced, this problem can be solved more easily than the full problem. Hence, in degeneration, our main interest is to find conditions under which the full problem tends to the degenerate problem, because then we may concentrate on the lower-order problem. A theorem due to Tickhonov [21] relates the solutions of the two problems, under the restriction of various assumptions.

(ii) *Asymptotic Expansion*: To obtain the full solution of the singularly perturbed system, suppose the solution has the form

$$\begin{aligned} x_1(t,\epsilon) &= x_{1_o}(t,\epsilon) + x_{1_c}(\tau,\epsilon) \\ x_2(t,\epsilon) &= x_{2_o}(t,\epsilon) + x_{2_c}(\tau,\epsilon) \end{aligned}$$

where $x_{1_o}(t,\epsilon)$ and $x_{2_o}(t,\epsilon)$ (the outer solutions), and $x_{1_c}(\tau,\epsilon)$ and $x_{2_c}(\tau,\epsilon)$ (the boundary layer correction solutions), are all assumed to have asymptotic expansions. The outer solutions are close to the exact solutions ($x_1(t,\epsilon)$ and $x_2(t,\epsilon)$, respectively) for values away from $t = t_0$ or outside the boundary layer, while the boundary layer corrections are significant only near $t = t_0$ or inside the boundary layer. Thus we ask that all the terms in the expansions for $x_{1_c}(\tau,\epsilon)$ and $x_{2_c}(\tau,\epsilon)$ tend to zero as $\tau \to \infty$. By computing these asymptotic expansions it is possible to separate the analysis of the singularly perturbed system into two parts: fast and slow. The complete details about when it is possible to compute the expansions and how to actually obtain them can be found in the literature [21].

(iii) *Time Scale Decomposition*: Often one is interested in controlling a singularly perturbed system. One approach to this is to decompose the problem into two control problems: one

for the fast subsystem and one for the slow subsystem. In this case the external input might be a state feedback control law of the form

$$u = K_s x_1 + K_f x_2$$

where K_s and K_f are the gain matrices for the slow and fast subsystems, respectively. The highlight of this approach is that it is often computationally much simpler to obtain the composite controls from the lower-order subsystems than to obtain a completely coupled feedback control for the original higher-order system. We believe that it is this aspect of SPATS that is particularly applicable to the design of convergent learning algorithms in neural networks.

B. Hopfield Networks with Learning

Next consider the Hopfield-type recursive neural network shown in Figure 9. We assume

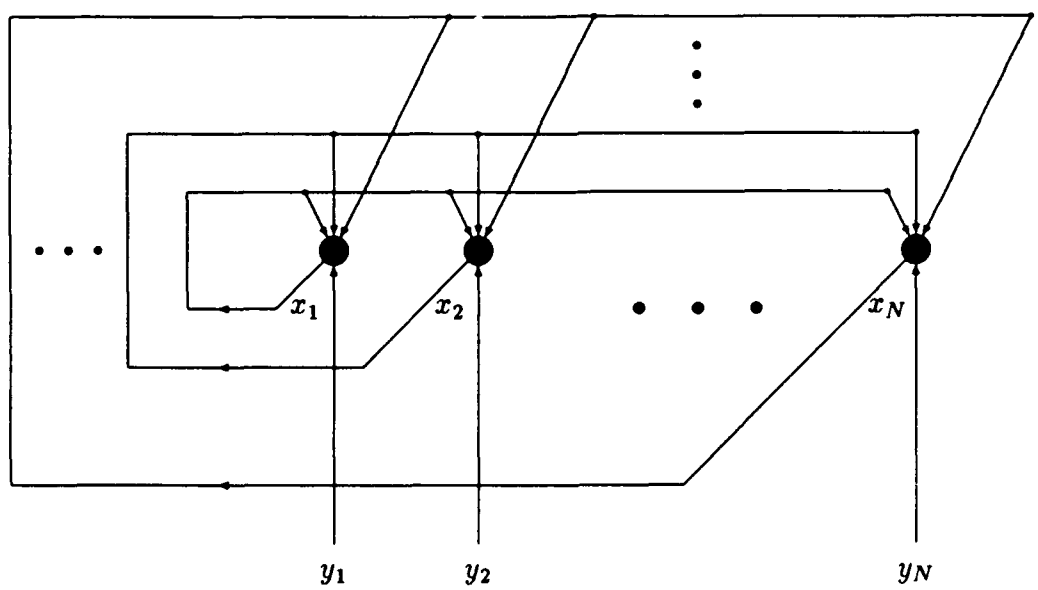


Figure 9: Recursive neural network.

that the dynamic behavior of this interacting system of neurons is governed by the following set of nonlinear differential equations [22].

$$C_i \frac{dx_i}{dt} = -\frac{x_i}{R_i} + \sum_{j=1}^{N} w_{ij} h(x_j) + y_i$$

for $i = 1, \cdots, N$. In this equation we identify

N: number of neurons.

x: output potential of a neuron.

C: cell capacitance.

R: membrane resistance.

w: synaptic weight.

$h(x)$: sigmoidal function describing the firing rate.

y: external input to the net.

(i) *Computing*: Typically, such a system is used in an operational or "computing" mode, with the weight vectors fixed to give the network some desired property. For example, the Hopfield network might be used as an associative memory. Because the network is described by a system of nonlinear differential equations, it contains a number of equilibrium points. When run in an operational mode the network receives an input vector which acts as a forcing function (or initial condition) for the system equations. Eventually the system's trajectory converges to an equilibrium, which defines the output pattern. Because the equilibria of such nonlinear systems usually have a non-trivial basin of attraction it is possible to converge to the same equilibrium point if we input a vector which is different from the original input but inside the region of attraction [23]. In this way the network acts as a memory, associating a particular output pattern with a particular input pattern. This feature leads to obvious applications for pattern recognition problems in image and signal processing. Other applications of Hopfield nets are in the area of optimization. In this case the weights are preset to values that are determined by the optimization problem one wishes to solve. If this is done properly, the equilibrium state of the net will define the solution to the problem [16].

A key question in the use of Hopfield nets is how to pick the weights so the net solves a given optimization problem or stores a desired equilibrium point. One answer to the question has been given in [24]. In this work it is shown how to design the weights *a priori* so that the system has a desired equilibria set. The solution involves solving a linear programming problem (which can be done by a second net).

(ii) *Learning*: An alternate approach to selecting the weights in a Hopfield net to produce a desired equilibrium point is to introduce learning. That is, we design an algorithm which adjusts the weights online until the network output is satisfactory. We call this a "learning" mode of operation for the Hopfield neural network and will refer to the "Hopfield net with learning" or "the learning Hopfield network." Note that learning is not a feature in most analyses of the Hopfield network, although it commonly arises in applications of feedforward neural networks.

The next obvious question would be what learning algorithm should be used. For the continuous Hopfield network described above, a possible weight adjustment algorithm might

be,

$$\lambda \frac{dw_{ij}}{dt} = x_j[x_i - \sum_{k=1}^{N} w_{ik}x_k]e_i.$$

Here λ could be thought of as the time constant of the change of the synaptic weight. $e_i = (x_i^d - x_i)$ is the error between the desired equilibria and the actual equilibria. λ is also called the learning gain and is chosen so that the learning proceeds at a slower rate than the time constant of the neural net. This is a logical extension of the idea of learning in feedforward nets. We see the learning as a loop operating around the input/output behavior of the net. As in feedforward nets, choosing the learning gain would likely be one of the most difficult tasks in developing a Hopfield neural network which can learn. Further, a more fundamental problem is that of determining the best structure for a learning algorithm

$$\lambda \frac{dw}{dt} = f(w, x, e)$$

that can ensure convergence of the learning process, while maintaining overall system stability. The problem is hard because of the highly coupled nonlinearities inherent in the process.

C. An Example of SPATS in a Learning Hopfield Net

(i) *Interpretation of Computing and Learning*: Consider a Hopfield net with learning as described in the previous subsection. We can interpret the computing mode of operation defined by the system equations as a "fast" synaptic event associated with a small time constant (RC), whereas the learning (synaptic weight change) can be thought of as a "slow" process associated with a large time constant (λ). Thus the computing and learning processes together constitute a "time-scale" behavior in the recursive neural network. This makes it natural to consider applying SPATS techniques to the convergence analysis and design of recursive neural networks with learning.

In order to fit the learning and computing equations into the standard framework of SPATS we recast the net equations as

$$\begin{aligned} \frac{dw}{dt} &= f(w, x, e) \\ \epsilon \frac{dx}{dt} &= g(w, x, y) \end{aligned}$$

with the initial conditions $x(0)$ and $w(0)$. Here the neuron states x (representing the computing process) and the network weights w (representing the learning process) have been redefined as fast and slow variables of order n and n^2, respectively. ϵ is a small positive parameter representing the time constant associated with the fast state x, e is the error vector, and y is the input vector of the net. Note that we have absorbed λ into $f(\cdot)$. Also

note that ϵ is in some sense an artificial constant that can be chosen by the designer. The key point is that learning should proceed slower than computing. Hence there should be a singularly perturbed character inherent in the process. This is introduced via the parameter ϵ. We believe that this set of equations, called the singularly perturbed neural network equations, will exhibit all the characteristic features of singularly perturbed systems and, consequently, we expect to be able to take advantage of this in the analysis and design of ANNs and ANN-based systems.

(ii) *Example*: To illustrate the fact that the Hopfield net with learning does in fact exhibit a singularly perturbed character, consider a two neuron Hopfield-net. Let the net be described by the computing equations

$$\begin{aligned} \epsilon\dot{x}_1 &= -x_1 + w_{11}h(x_1) + w_{12}h(x_2) + y_1 \\ \epsilon\dot{x}_2 &= -x_2 + w_{21}h(x_1) + w_{22}h(x_2) + y_2 \end{aligned}$$

and let the weight update equations be

$$\begin{aligned} \dot{w}_{11} &= x_1 e_1 \\ \dot{w}_{12} &= x_1 e_2 \\ \dot{w}_{21} &= x_2 e_1 \\ \dot{w}_{22} &= x_2 e_2 \end{aligned}$$

Here the error e_i is defined by

$$e_i = x_i^d - x_i$$

with x_i^d the desired output state for the ith neuron, in response to the input signal y_i. The weight update equations are simpler than the general form suggested above. They were constructed to be similar in structure to the Delta rule used for training perceptrons.

Figure 10 shows a typical trajectory for this set of equations when driven with the input/output pair $y = (1,1)$, $x^d = (0,0)$. It can be seen that the system approaches the desired equilibrium state. We found this to be true for all the training pairs (y, x^d) that we tried. This is interesting because we did not choose the weights *a priori* to ensure convergence to the correct output. Rather, the system learned the correct weight values on its own. Further, define $\bar{w}_{ij}$ to be the steady-state value of the weights w_{ij} after the complete system converges. Then let the system be operated independent of the weight adjustment equations, using the $\bar{w}_{ij}$ as constant weights. In this case, the output of the net will still converge to the correct values. This suggests that such an approach to learning can be effective. A final comment on the example system is indicated in Figure 11. Here it is seen that the system exhibits the classic SPATS characteristic of multiple time scales. We see the initial region of rapid transition, corresponding to the fast dynamics of the computing portion of the net

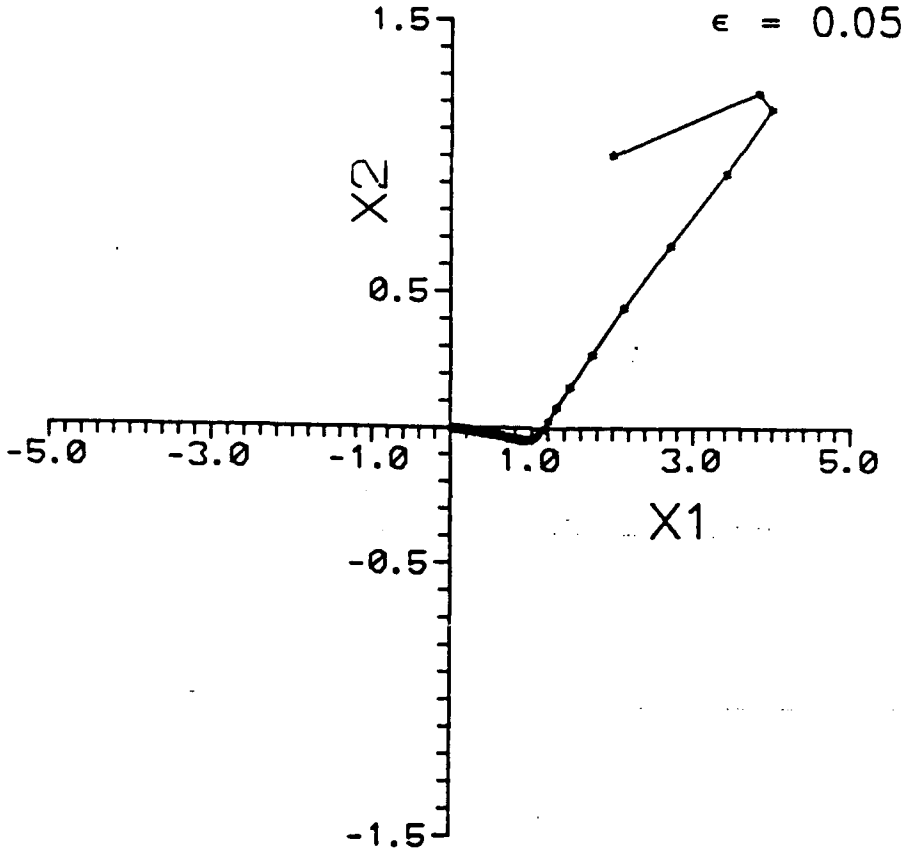


Figure 10: State trajectory of *example* system.

and we also see the long time constant of the learning equations. Thus the Hopfield net with learning is in fact a singularly perturbed system. Future research activities, as described below, will consider how to exploit this fact for analysis and design of neural networks.

VII. Future Research Directions

The investigations leading to the results described in the previous five sections also led to new questions. In this final section we briefly describe three of these questions. We currently have research proposals pending with both NSF and AFOSR to continue our investigations in each of these areas.

A. Learning in Feedforward Neural Nets

One problem we have encountered in our research deals with learning and generalization in neural nets. In most experiments we must use random training pairs to ensure adequate coverage of the input/output space. However, there is no good basis for such an approach.

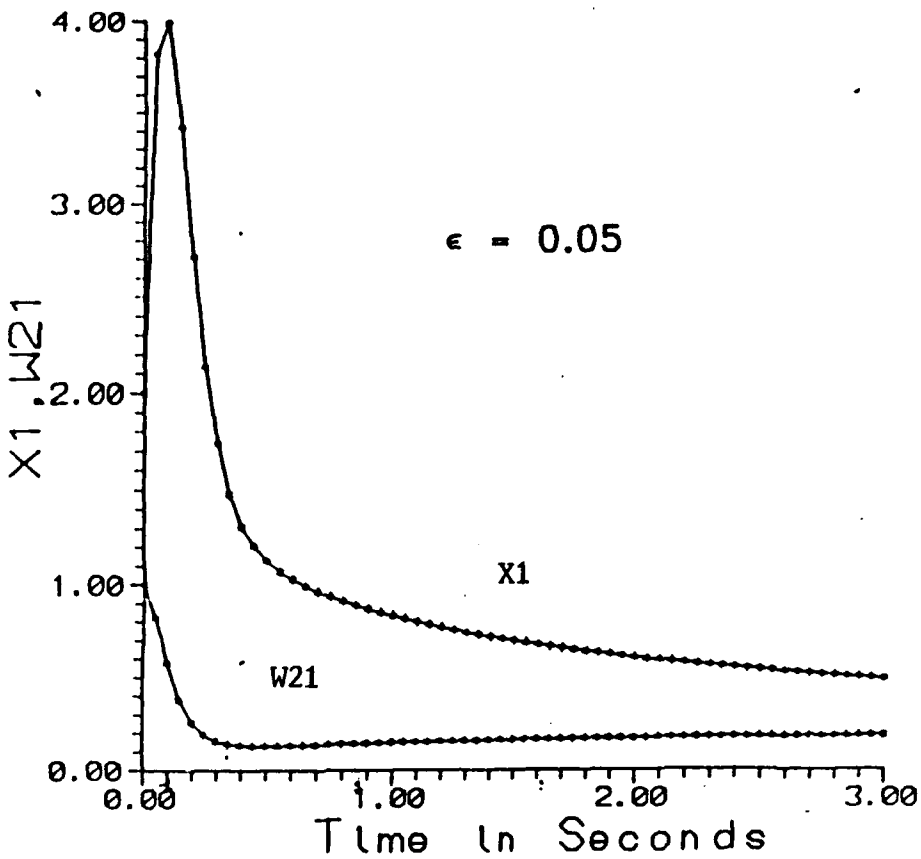


Figure 11: State x_1 and weight w_{21} of the example system.

We have developed some preliminary ideas for overcoming this problem by viewing it from a control-theoretic perspective. We discuss two ideas below.

(i) *Persistence of Excitation for Neural Nets*: Suppose we are given a discrete-time, nonlinear system defined by the input-output map

$$x_{k+1} = f(x_k, u_k)$$

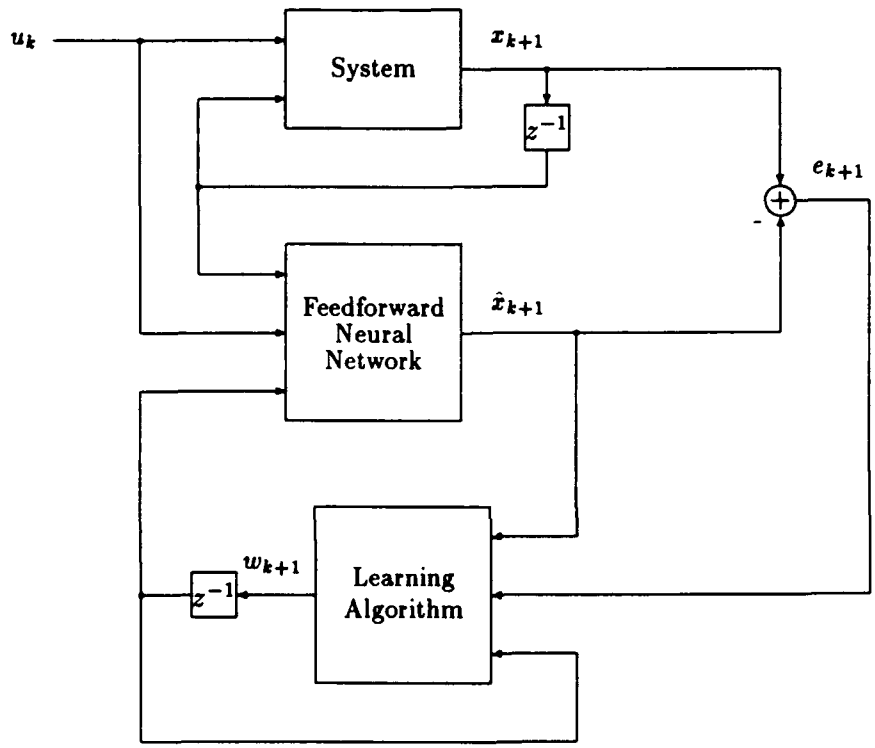
where x_k is the output of the system and u_k is the input to the system. Also suppose we have a feedforward neural net defined by

$$\hat{x}_{k+1} = \hat{f}(x_k, u_k, w_k)$$

where $\hat{x}_k$ is the output of the neural network and w_k is a vector of the weights of the net at time $t = k$. It is assumed that the input signal u_k is available to the net. Further, we suppose that the weights are updated according to the algorithm

$$w_{k+1} = g(\hat{x}_k, w_k, e_k)$$

where $e_k = x_k - \hat{x}_k$. This is depicted in Figure 12. Note that the learning algorithm does not necessarily need to be backpropagation, but does represent a dynamical system with respect to the weights (just as backpropagation does). This interpretation is somewhat limited in that it precludes consideration of some type of learning schemes, such as ones based on competitive learning [25], clustering [26], or random mappings within the net architecture [27,2,28]. However, it is an accurate depiction of many learning algorithms.



System: $x_{k+1} = f(x_k, u_k)$

Net: $\hat{x}_{k+1} = \hat{f}(x_k, u_k, w_k)$

Learning: $w_{k+1} = g(\hat{x}_k, w_k, e_k)$

Figure 12: Learning as a nonlinear control problem.

The problem we are interested in is this: for the configuration shown in Figure 12, with g fixed, how should the training inputs u_k be chosen to ensure that the net will learn the training pairs and will also give the correct output for inputs it has never seen? Typically this is done by choosing u_k randomly, according to some distribution. An alternate approach is to introduce the issue of persistence of excitation.

The problem of persistence of excitation has been studied extensively in the adaptive

control community. The following example illustrates the concept [29]. Consider a linear plant with known structure, but unknown parameters

$$\dot{y} = -a_p y + b_p u$$

The task is to find a feedback control input u so that the output of the closed-loop system tracks the reference model

$$\dot{y}_m = a_m y_m + b_m r(t)$$

In [29] the following control law is suggested. Let

$$u(t) = \hat{a}_y(t)y(t) + \hat{a}_r(t)r(t)$$

where the parameters $\hat{a}_y(t)$ and $\hat{a}_r(t)$ are adjusted according to the adaptation law

$$\dot{\hat{a}}_y = -\text{sgn}(b_p)\gamma e y$$

$$\dot{\hat{a}}_r = -\text{sgn}(b_p)\gamma e r$$

Here $e = y - y_m$ is the tracking error, γ is an adaptation gain, and the sign of b_p must be known. It can be shown that with this control law and adaptation algorithm then

$$\lim_{t\to\infty} e(t) = 0$$

If we examine this control system we see that the ideal choice of the controller gains is $a_y^* = (a_p - a_m)/b_p$ and $a_r^* = b_m/b_p$. A logical question at this point would be to ask if the parameters converge to the optimal value when the error converges to zero. The (surprising) answer is sometimes. It turns out that it is possible to have the error converge to zero even when the controller parameters converge to some value other than the optimal gains. For the example above this can occur when the reference input $r(t)$ is a simple step input. However, if the input to the system is a sinusoidal signal, then it can be shown that the parameters converge to their optimal values as the tracking error goes to zero. We describe this by saying that the sinusoidal input is persistently exciting for the first-order system, but the step input is not. In general, it turns out that to identify $2n$ parameters in a linear system (of order n) it is necessary to include n sinusoids of different frequency in the reference signal. (We must know the order of the system and the sign of the coefficient of the highest-order derivative of the forcing function of the plant.)

It is possible to extend these ideas to some classes of nonlinear systems. For instance, in [29] it is shown how to achieve adaptive control, with parameter convergence of the system

$$\dot{y} = -a_p y - c_p f(y) + b_p u$$

where $f(\cdot)$ is a known nonlinear function and a_p, b_p, and c_p are unknown parameters. It is also shown in [29] how to develop adaptive controllers (based on sliding-mode control

principles) for systems of the form

$$y^{(n)} + \sum_{i=1}^{n} \alpha_i f_i(x,t) = bu$$

where $x = (y, \dot{y}, \ldots y^{(n-1)})^T$ is the state vector, the $f_i(\cdot)$'s are known nonlinear functions and the α_i's and b are unknown parameters. Unfortunately, even for these relatively tame systems it is only possible to give very general conditions for parameter convergence and persistence of excitation. For instance, it is not yet possible to give a nice result on the number of sinusoids required in the input signal.

Now consider the implications these observations have for the problem of learning in neural nets. First, although not apparent from the brief comments above, it is likely that any successful learning scheme will involve feedback of the modelling error into the input signal u_k shown in Figure 12. This is true in the linear case and in the nonlinear examples given in [29]. However, our original interpretation of learning in a neural net shown in Figure 12 did not consider such feedback, but rather assumed an arbitrary plant input signal for training. This suggests that our learning algorithm should affect not only the weight update, but also the plant input. That is, we should control the plant at the same time we are identifying it.

Second, even with feedback in place, it will be necessary that we have some idea about what constitutes a persistently exciting input signal for learning with neural nets. Because this question does not have a satisfactory answer for even the simple nonlinear example given above (where the unknown weights are linearly related to the nonlinear functions, as opposed to the neural net case where the unknown weights are embedded in the nonlinearity), it will likely be hard to find a satisfactory answer for the neural net case. This suggests the need for more basic research on the topic of nonlinear adaptive control.

A third comment is that if our objective is to control a system using a neural network it may not be essential that we achieve convergence of the weights to the exact values of the optimal net which models a system or its inverse. If the control objective is achieved, then it may not be important that the parameters have converged. However, from the perspective of learning, our main interest is when the weights of the net converge to values that are correct in the sense of providing the best possible representation of the desired nonlinear mapping over the desired region of the input space.

(ii) *A Stochastic Approach to Learning*: Searching for a persistently exciting input for a system is a deterministic approach to the learning problem. The goal is to find a deterministic sequence of inputs that will excite the system to produce a type of basis for the output space, so we can ensure adequate representation of the input-output space. An alternate approach is to obtain the necessary coverage of the input-output space through the use of random input signals. This is the approach used in almost every example of training given in the

literature. One shortcoming of this approach when used with most training algorithms, such as backpropagation, is that the training algorithm does not take into account the fact that the input sequence is random. In this section we discuss an approach to stochastic training based on optimal filter theory (i.e., Kalman filtering) which does take into account the randomness of the input sequence.

The results we describe were first reported by Wabgaonkar and Stubberub [30]. An implicit assumption in this work is that there exists a network that will solve the problem and that we know the number of neurons in its hidden layer. In practice this may not be true. However, a practical learning system based on this approach would probably assume some number of neurons, perform the estimation, increase the number, and iterate until the error is reduced to an acceptable level. Thus this is a reasonable assumption and is implicit in many other papers on neural nets (in the Narendra paper [8], for example).

We briefly summarize the Kalman filter [31]. Consider a dynamical system defined by

$$\begin{aligned} x_{k+1} &= Ax_k + w_k \\ z_k &= Hx_k + v_k \end{aligned}$$

Here z_k is our output measurement, the system matrices A and H are known, and w_k and v_k are uncorrelated white-noise sequences representing input noise and measurement noise, respectively, with known covariance matrices. The Kalman filtering problem is to find an estimate $\hat{x}_k$ of the system state x_k such that the expected value of the mean-square error $e_k = x_k - \hat{x}_k$ is minimized. The solution to this problem is well-known and is given by

$$\hat{x}_k = \hat{x}_k + K_k(z_k - H\hat{x}_k)$$

where K_k is a time-varying gain matrix obtained by solving a Riccati equation. Note that this problem statement was for a linear system. For nonlinear systems a similar solution to the state estimation problem can be developed. This is called an extended Kalman filter and involves a linearization of the nonlinear system.

Now, suppose that we wish to learn an input-output mapping $y_k = f(u_k)$. As mentioned earlier, we assume that a solution exists using a standard feedforward net (one hidden layer with a fixed number of neurons) with nonlinearities in the neurons defined by

$$g^j(x) = a_0^j + b_1^j \cos(2\pi b_2^j x) + a_1^j \sin(2\pi a_2^j x)$$

Notice that the form of the assumed nonlinearity causes the net to perform a type of Fourier series approximation of the unknown function. Next define a vector x composed of all the weights and biases, together with all the parameters $a_0^j, b_1^j, b_2^j, a_1^j$, and a_2^j. If a real net exists that implements the nonlinear function of interest, then the vector x of the real net will be a constant so the net could be described by

$$\begin{aligned} x_{k+1} &= x_k \\ d_k &= y_k + v_k = f_k(x_k, u_k) + v_k \end{aligned}$$

where x is the state of the true net, y is the actual output, u is the actual input, and v_k is viewed as random measurement noise (which has some assumed statistics) which corrupts the actual output, giving the desired or measured output d_k.

Given this model of the unknown function it is straightforward to derive an extended Kalman filter for the system. Simulation results arising from such an analysis indicate that the method works quite well [30]. However, these results considered only one- and two-neuron systems. It will be important to see how well the method will work for systems with a larger number of neurons. Also, it is not clear that the use of sinusoidal nonlinearities is optimal. More work must be done to determine the best choice of neuron transfer functions. The important thing to note is that by using a Kalman filter to estimate the parameters of the net (i.e., the weights) we are able to develop a training algorithm which takes into consideration the random nature of the input during training.

(iii) *Proposed Research Activities* In light of the discussion presented above, the following research questions are posed on two specific problems.

1. Persistence of Excitation and Feedback in Learning: We propose to study the development of learning algorithms for feedforward networks using traditional adaptive control techniques as a point of departure. We consider two different tracks of investigation.

   (a) *Persistence of excitation*: We propose to study the role of persistence of excitation in the generation of training data. We first restrict ourselves to feedforward nets with one hidden layer and sigmoidal nonlinearities. We assume that such a net accurately describes an unknown nonlinear dynamical system for some set of optimal weights. We then ask the question: What is the nature of a persistently exciting input for learning in a scheme such as in Figure 12 so the weights will converge to the optimal weights? We will attempt to answer this question in as general a way as possible, but expect that the answers will have to be given for specific classes of functions to be learned.

   (b) *Feedback*: As noted above, most adaptive control schemes use feedback of the identification error or the tracking error into the input. This is unique in the learning algorithms typically used for training feedforward nets. We propose to investigate the use of feedback as an aid to training.

2. Kalman Filtering Approach to Learning: We propose to study in more depth the optimal filtering approach of [30]. Specifically we will consider the choice of nonlinearity to use in the net. We will consider this from the perspective of the representation problem from approximation theory. In particular, we will consider a marriage of the Kalman filtering approach to learning to the radial basis function approach to the representation problem.

B. Learning in Recursive Neural Nets

The discussion in section VI regarding singular perturbations and time scales applied to neural nets suggests the following research questions.

1. SPATS Analysis of Neural Networks: We have seen how a Hopfield net with learning could be interpreted as a singularly perturbed system and we presented an example to show that such a system does in fact produce multiple time-scale behavior. The first step is to perform a complete analysis of the general N-neuron Hopfield net with learning from the perspective of SPATS. In particular we will demonstrate that the Hopfield net with learning satisfies the assumptions necessary to compute both degenerate solutions and asymptotic expansions for the system. We will also compute such solutions and expansions for representative examples. It is expected that such an analysis will offer insight into how to ensure convergence of the net with learning. Note that the analysis requires that we assume a structure for the learning algorithm. Initially, we will begin with the Delta rule-type equation given in the example above. However, as the research progresses we expect to gain insight into the type of learning law that will be most effective. This insight can be gained by first applying SPATS analysis techniques to the preliminary learning algorithm. This also leads to the following problem. Suppose we are given a system

$$\epsilon \frac{dx}{dt} = f(w, x, y)$$

The question is what function $g(\cdot, \cdot, \cdot)$ satisfying

$$\frac{dw}{dt} = g(w, x, e)$$

will cause $e \to 0$ (or some other criterion)? This problem is an inverse singular perturbation problem, and is novel in the study of SPATS. However, this problem arises in the design of learning algorithms for a Hopfield net. We propose to address this problem as part of our future research activities.

2. SPAT-Based Design of Learning in Neural Networks: One feature of SPATS is that it is often possible to decouple the fast and slow modes of the singularly perturbed system for the purpose of analysis and design. We will attempt to do this for the Hopfield network with learning. We can then design stabilizing controls separately for both the computing (fast) modes and the learning (slow) modes. If this is successful it may be possible to design learning algorithms that guarantee convergence without the usual problems of local minima. At the least, we should be able to develop sufficient conditions for such convergence.

3. Systems Containing Neural Networks: The third goal of the research is to consider the design of neural network controllers for dynamical systems. Consider Figure 13, where we suppose that the controller C in Figure 13 is a continuous Hopfield network with learning. In this case the closed-loop system will also exhibit a singularly perturbed character. There are two special cases that can arise.

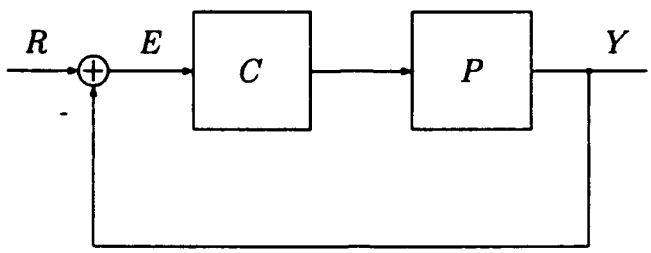


Figure 13: Unity feedback control system.

(a) *Normal Plant*: In this case we suppose that our plant is a linear or nonlinear system. If we also suppose that the neural net is learning at some rate that is significantly slower than the closed-loop dynamics, then we can view the closed loop system as a singularly perturbed system. We can then decompose the entire closed-loop system equations into their fast and slow modes.

(b) *Singularly Perturbed Plant*: In this case we will have a closed-loop system which exhibits time scale behavior. This is true for two sub-classes: (1) the neural net is used in an operational mode and (2) the neural net is used in a learning mode. In either case the closed-loop system will have both fast and slow modes.

We propose to use SPATS to specify how to design both the learning and computational parts of a neural network to ensure learning convergence and closed-loop stability for the different cases enumerate above.

C. Kalman Filtering and Real-Time Adaptive LQR Control

In future work we expect to develop the application of neural nets to solving optimal control problems in more detail. In particular, the following topics are currently under study.

1. Further investigation of the structure of the neural net. We are interested in the fact that the neural net will be sparsely interconnected for distant-horizon problems due to the structural issues we mentioned above.

2. As noted previously, we are developing numerical examples using P-SPICE to illustrate the technique and we will compare these examples to conventional solutions of optimal control problems.

3. We hope to apply the method to solving general Riccati equation problems. This will allow us to extend the method from the LQR problem to other quadratic cost problems such as LQG and Kalman filtering. The following logic leads us to believe this is possible. We have shown that a neural net can be used to solve an LQR optimal control problem. However, the conventional approach to solving such a problem involves the solution of a Riccati equation. Thus we can say that the neural net is solving a Riccati equation. Now consider a Kalman filtering problem. It is well known that this problem is also solved by solving a Riccati equation. Hence we propose the following approach to solving the Kalman filtering problem. First, map from the Kalman filtering problem to its associated Riccati equation. Then map from the Riccati equation to its associated optimal control problem. Then solve the optimal control problem using a neural net. Finally, map back from the neural net solution of the optimal control problem, to the Riccati equation solution, and then to the Kalman filtering solution. This approach is currently being considered by an unfunded graduate student.

4. Finally, we are interested in using the technique in real-time adaptive control schemes. In [7] we showed how to identify the state-space model of a linear system using a neural net. We can use this, together with the optimal control result given above, to suggest the configuration shown in Figure 14. As shown, the identification net passes the model to the optimal control net, which computes the optimal feedback gains. We are interested in studying the closed-loop stability and convergence properties of this scheme.

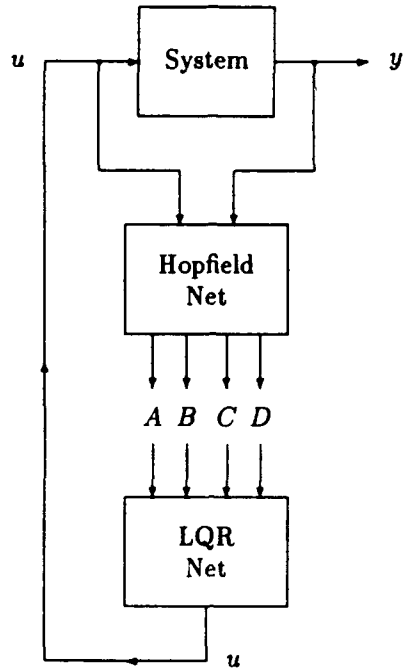


Figure 14: Real-time, adaptive control using neural networks.

References

[1] K. Moore, "Neural networks for guidance, navigation, and control of exoatmospheric interceptors," Final Report, 1990 Summer Faculty Fellowship Program, AFATL/SAI, Air Force Armament Laboratory, Eglin AFB, Florida, September 1990.

[2] K. Moore, "A reinforcement-learning neural network for the control of nonlinear systems," in *Proceedings of 1991 American Control Conference*, (Boston, Massachusetts), June 1991.

[3] K. Moore and S. Naidu, "Linear quadratic regulation using neural networks," in *Proceedings of the 1991 International Joint Conference on Neural Networks*, (Seattle, Washington), August 1991.

[4] K. Moore and S. Naidu, "Optimal control using neural networks," in *Proceedings of AI-91*, (Jackson, Wyoming), September 1991.

[5] M. Waddoups and K. Moore, "Neural networks for iterative learning control," in *Proceedings of 1992 American Control Conference*, (Chicago, Illinois), June 1992.

[6] K. Moore and S. Naidu, "Singular perturbations and time scales in neural networks," in *Proceedings of* 30th *IEEE Conference on Decision and Control*, (Brighton, United Kingdom), December 1991.

[7] C. Zhang and K. Moore, "System identification using neural networks," in *Proceedings of* 30th *IEEE Conference on Decision and Control*, (Brighton, United Kingdom), December 1991.

[8] K. Narendra and K. Parthasarathy, "Identification and control of dynamical systems using neural networks," *IEEE Transactions on Neural Networks*, vol. 1, pp. 4–27, March 1990.

[9] J. Craig, *Adaptive Control of Robot Manipulators*. Reading, Massachusetts: Addison-Wesley, 1988.

[10] K. Narendra and M. Thathachar, "Learning automata — a survey," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 4, pp. 323–334, July 1974.

[11] A. Klopf, *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Washington, D.C.: Hemisphere, 1982.

[12] A. Barto, R. Sutton, and C. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, pp. 834–846, September 1983.

[13] J. A. Franklin, "Refinement of robot motor skills through reinforcement learning," in *Proceedings of* 27th *Conference on Decision and Control*, (Austin, Texas), pp. 1096–1101, December 1988.

[14] C. W. Anderson, "Strategy learning with multilayer connectionist representations," GTE Labs Technical Report TR87-509.3, GTE Laboratories, Waltham, Massachusetts, May 1988.

[15] R. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9–44, 1988.

[16] M. Kennedy and L. Chua, "Neural networks for nonlinear programming," *IEEE Transactions on Circuits and Systems*, vol. 35, pp. 554–562, May 1988.

[17] T. Kailath, *Linear Systems*. Englewood Cliffs, New Jersey: Prentice-Hall, 1980.

[18] H. Taha, *Operations Research: An Introduction*. New York, New York: MacMillan Publishers, 1982.

[19] S. Arimoto, S. Kawamura, and F. Miyazaki, "Bettering operation of robots by learning," *Journal of Robotic Systems*, vol. 1, pp. 123–140, March 1984.

[20] K. Moore, *Design Techniques for Transient Response Control.* PhD thesis, Texas A&M University, College Station, Texas, 1989.

[21] D. Naidu, *Singular Perturbation Methodology in Control Systems, IEE Control Engineering Series.* London, United Kingdom: Peter Peregrinus, LTd., 1988.

[22] J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proceedings of the National Academy of Science*, vol. 81, pp. 3088–3092, May 1984.

[23] M. Cohen and S. Grossberg, "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks," *IEEE Transactions on Systems, Man, and Cybernetic*, vol. 13, pp. 815–826, September/October 1983.

[24] A.Guez, V. Protopopsecu, and J. Barhen, "On the stability, storage capacity, and design of nonlinear continuous neural networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, pp. 80–87, January/February 1988.

[25] D. Rumelhart and J. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vols. I and II.* Cambridge, Massachusetts: MIT Press, 1986.

[26] N. Weymaere and J. Martens, "A fast and robust learning algorithm for feedforward neural networks," *Neural Networks*, vol. 4, pp. 361–369, 1991.

[27] W. M. III, F. Glanz, and L. K. III, "CMAC: an associative neural network alternative to backpropagation," *Proceedings of the IEEE*, vol. 78, pp. 1561–1567, October 1990.

[28] S. Gallant, "A connection learning algorithm with provable generalization and scaling bounds," *Neural Networks*, vol. 3, pp. 191–201, 1990.

[29] J. Slotine and W. Li, *Applied Nonlinear Control.* Englewood Cliffs, New Jersey: Prentice-Hall, 1991.

[30] H. Wabgaonkar and A. Stubberud, "Approximation and estimation techniques for neural networks," in *Proceedings of the 29th IEEE Conference on Decision and Control*, (Honolulu, Hawaii), pp. 2736–2740, December 1990.

[31] B. Anderson and J. Moore, *Optimal Filtering.* Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

FINAL REPORT

Fire Control System for a Laser Aimed Machine Gun

by

William D. Siuru, Jr., PhD, PE

Senior Research Associate
Space and Flight Systems Laboratory
University of Colorado at Colorado Springs

December 15, 1991

USAF-UES MINIGRANT PROGRAM

Sponsored by

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. Universal Energy Systems' efficient administration of the program was greatly appreciated.

Special appreciation is given to Gary Cunning for his programming of many algorithms used in this investigation and for his research in optimizing filter gains and data sampling rates.

Excellent administrative support was provided by Karen Clevenger. Brian Siuru proofread the final report.

ABSTRACT

Techniques were investigated to modify the BSTING fire control system to make it more responsive and provide continuously updated aiming corrections. The ultimate goal was a "point and shoot" capability while still retaining a simple, gun-mounted stand-alone system. A correction scheme was devised that accounted for the fact that azimuth and elevation rates are measured along gun axis rather than the required line-of-sight axis. Also investigated was the error that results if this correction is not included. Another technique investigated locked the gun to the laser beam during the first two seconds while the initial solution is computed and then the gun was unlocked to allow continously updated solutions. An investigation was made to see how system response could be improved by using adaptive Kalman filtering and increasing the data sampling rates. An analysis was performed to determine the sensitivity of aiming corrections to the accuracy of the sensor measurements. A brief analysis showed the conditions where elevation angle must be included in the FCS algorithms and the "small angle" approximation is not appropriate. A new fire control system using externally provided measurements was designed and evaluated.

CHAPTER I

INTRODUCTION

Objectives and Approach

While the current BSTING system has shown significant promise, its greatest limitation is that it requires several seconds for measuring parameters, performing computations, implementing the aiming corrections and relocating the laser spot on the target before the weapon is ready to fire. Also, once the solution is implemented, the process must be repeated before the gun can be realigned with respect to the laser beam. This results mainly because while the laser is used to measure the range and range rate, gun elevation and azimuth rates are measured by the gun itself, requiring the gun to be locked to the laser beam while measurements are made.

Ideally, it would be desirable to make measurements, compute corrections and execute them continuously. In other words, the system would provide almost a "point and shoot" capability.

OBJECTIVES

The purpose of this research was to determine if it is possible to design a gun-mounted, stand-alone, system that would allow continuous gun aiming solutions rather then require discrete time periods for tracking, computation and aiming. Also it is desirable that the system provide and implement aiming corrections in a minimum amount of time.

First, an investigation was done to see if the current BSTING algorithms could be modified to provide a continuous aiming solution. Secondly, a fire control system using externally provided measurements was designed and evaluated. Finally, an investigation was made to see how system response could be improved (i.e. capability approaching "point and shoot") through (a) use of adaptive Kalman filtering and (b) increasing data sampling rates.

APPROACH

As an initial step, an investigation was made to see if the current BSTING fire control algorithms could be theoretically modified to provide continuous tracking, computation and aiming. A more adaptable and capable BSTING-type system is desirable because of the inherent simplicity of

the system with its minimum of required measurements and completely "stand-alone" characteristics. This initial design incorporated a correction technique that compensated for the fact that in the BSTING system, azimuth and elevation rates are measured along the gun boresight rather than the laser beam line-of-sight as is required for most accurate corrections.

In order to reduce the number of parameters that had to be investigated and optimized, a brief investigation was done to determine the sensitivity of aiming corrections to the accuracy of the sensor measurements. Because the correction in the azimuth plane due to movement of the gun platform with respect to the target is most critical, and the largest source of error, this parameter was used in determining sensitivities.

The previous theoretical analysis showed it was possible to adapt the BSTING algorithms to provide continuously updated aiming corrections. However, this result was based on "unnoisy" data. Thus it was next necessary to include the noisy sensor measurements that results when the gunner is aiming the gun from a moving platform at a stationary target to see if the conclusions were valid under more realistic conditions. The noisy measurement data was simulated by generic noise profiles

superimposed on the theoretically required measurements. The noise profiles were based on data obtained during the BSTING flight test program (August-September 1990).

As mentioned previously, in the current BSTING system, the azimuth and elevation rates are measured along the gun boresight axis though computation of corrections are based on sensor measurements along the laser beam axis. To compensate for this difference, measurements are now made when the gun is locked to the laser beam axis. An investigation was made to determine the aiming error that results if the gun is "unlocked" from the laser beam and the azimuth and elevation rate sensors move with the gun when making measurements. Subsequently, this analysis was extended to include a technique where the gun was locked to the laser beam during the first two seconds while the initial solution is computed. This follows the technique used in the current BSTING system. After the initial computation was determined, then the gun was "unlocked" so that measurements could be used to provide a continuously updated solution using these computed aiming corrections as initial conditions.

The simple BSTING algorithms do not use the elevation angle in gravity correction. Thus an analysis was done to determine the magnitude of the error resulting from neglecting this measurement. The flight conditions where the elevation angle measurement should be included were identified.

An alternate fire control system was designed. This FCS uses external measurement of parameters in lieu of (or in addition to) sensors mounted on the gun itself as is the case with the current BSTING system. The goal was a more responsive fire control system that can provide "point and shoot" capability. Initially, this FCS was designed around ideal, that is unnoisy, sensor measurements. Subsequently, the analysis was extended using more realistic and noisy data. Again the noise profiles used were based on data obtained during the BSTING flight test program.

Finally, a rather rigorous analysis was conducted to determine if system response could be improved by using (a) optimization of the filter gains, (b) using adaptive Kalman filtering and (c) increased data sampling rates. These techniques were applied to the BSTING algorithms

CHAPTER II

MODIFICATIONS TO BSTING SYSTEM

Continuous Solution - Theoretical

As an initial step, an investigation was made to see if the current BSTING fire control algorithms could be modified to provide continuous tracking, computation and aiming. A more adaptable and capable BSTING-type system is desirable because of the inherent simplicity of the system with its minimum of required measurements and completely "stand-alone" character.

ANALYSIS OF BSTING AZIMUTH AIMING ERROR

Currently, the BSTING system makes range and azimuth/elevation measurements while the gun is "locked" to the laser beam. Thus the azimuth and elevation rates measured by the gun mounted sensor are the same as the desired LOS (laser) rate. However, if the two are "unlocked" (required to implement the correction) as would be required for a continuous computing system, the gun mounted rate sensors would not be measuring the

correct rates and therefore the resulting aiming corrections would be in error. This error would be directly related to the rate of change of the elevation and azimuth corrections.

The correction in the azimuth plane due to motion of the gun platform is much more critical to system accuracy than the rate corrections in the elevation plane, at least for straight and level flight maneuvers. In the BSTING system, the azimuth plane is the plane parallel to the gun boresight axis, whereas the elevation plane is normal to the ground. Thus the azimuth rate is directly related to the platform's horizontal velocity. Because azimuth rate errors are the greatest source of error, this parameter will be used in the subsequent analysis. Refer to figure II-1 for the correction in the azinmuth plane.

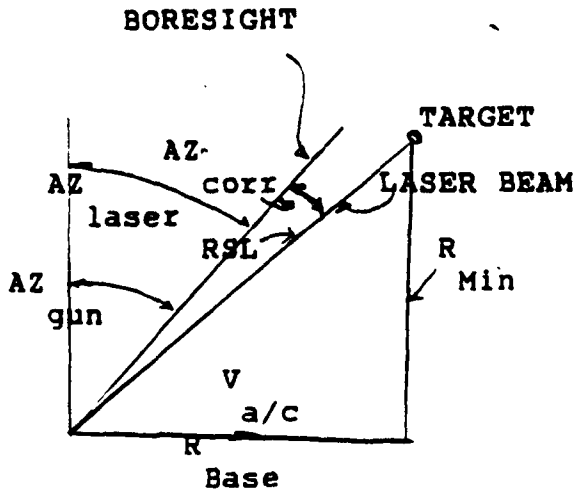


Figure II-1

Correction in Azimuth Plane
(Platform Moving at Velocity $V_{a/c}$ at Slant Range RSL)

From the figure

$$AZ_{gun} = AZ_{laser} + AZ_{corr} \qquad \text{(II-1)}$$

From the BSTING algorithms

$$AZ_{corr} = \frac{RSL}{V_{muz} \; Cos(EL_{corr})} \dot{AZ}_{laser} \qquad \text{(II-2)}$$

where: V_{muz} is the muzzle velocity

$\dot{AZ}_{laser}$ is the azimuth rate measured by the rate sensor along the centerline of the laser beam

EL_{corr} is the aiming correction in the elevation plane

Now when the rate sensor is mounted on the gun, the correction would be

$$AZ'_{corr} = \frac{RSL}{V_{muz} \; Cos(EL_{corr})} \dot{AZ}_{gun} \qquad \text{(II-3)}$$

For straight and level flight past a stationary target, the azimuth rate along the laser centerline can be determined using the following figure:

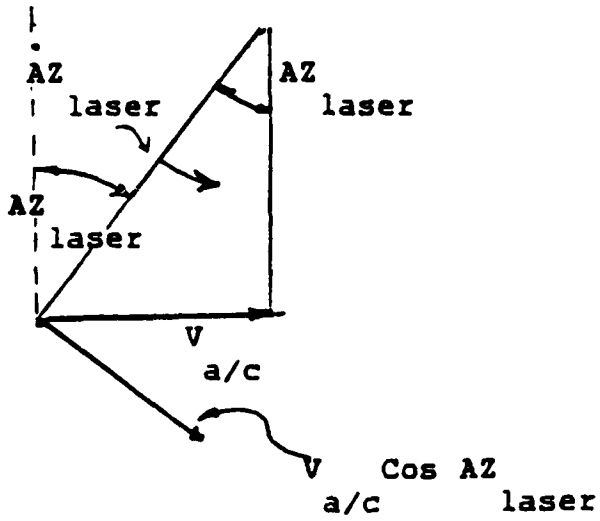


Figure II-2
Determination of Azimuth Rate in Laser Beam Plane

$$\dot{AZ}_{laser} = - \frac{V_{a/c} \cos AZ_{laser}}{RSL} = - \frac{V_{a/c} R_{min}}{RSL^2} \qquad \text{(II-4)}$$

If the azimuth correction were based on the gun mounted rate gyro measurement, as is the case for the current BSTING system, the correction would be

$$\dot{AZ}_{gun} = \frac{V_{a/c} \cos(AZ_{gun})}{RSL} = \frac{V_{a/c} \cos(AZ_{laser} + AZ_{corr})}{RSL} \qquad \text{(II-5)}$$

If $\cos(EL_{corr}) \cong 1$, which is the case for low altitudes

Then the error in correction is given by

$$AZ_{corr} - AZ'_{corr} = \frac{RSL}{V_{muz}} \{\dot{AZ}_{laser} - \dot{AZ}_{gun}\} \qquad \text{(II-6)}$$

$$= \frac{V_{a/c}}{V_{muz}} \{Cos(AZ_{laser}) - Cos(AZ_{laser} + AZ_{corr})\}$$

where AZ_{corr} is given by equation II-2

The error in azimuth aiming correction due to having the rate sensor mounted on the gun (BSTING system) rather than on the laser beam (required for maximum accuracy) axis was estimated using equation II-6 for several flight scenarios.

SCENARIOS

CASE	AIRCRAFT VELOCITY	MINIMUM SLANT RANGE	ALTITUDE ABOVE TARGET
1	75 kts	500 mtrs	250 ft
2	125 kts	1000 mtrs	250 ft
3	175 kts	1500 mtrs	250 ft
4	200 kts	2000 mtrs	250 ft

The results of this analysis are shown in Figure II-3.

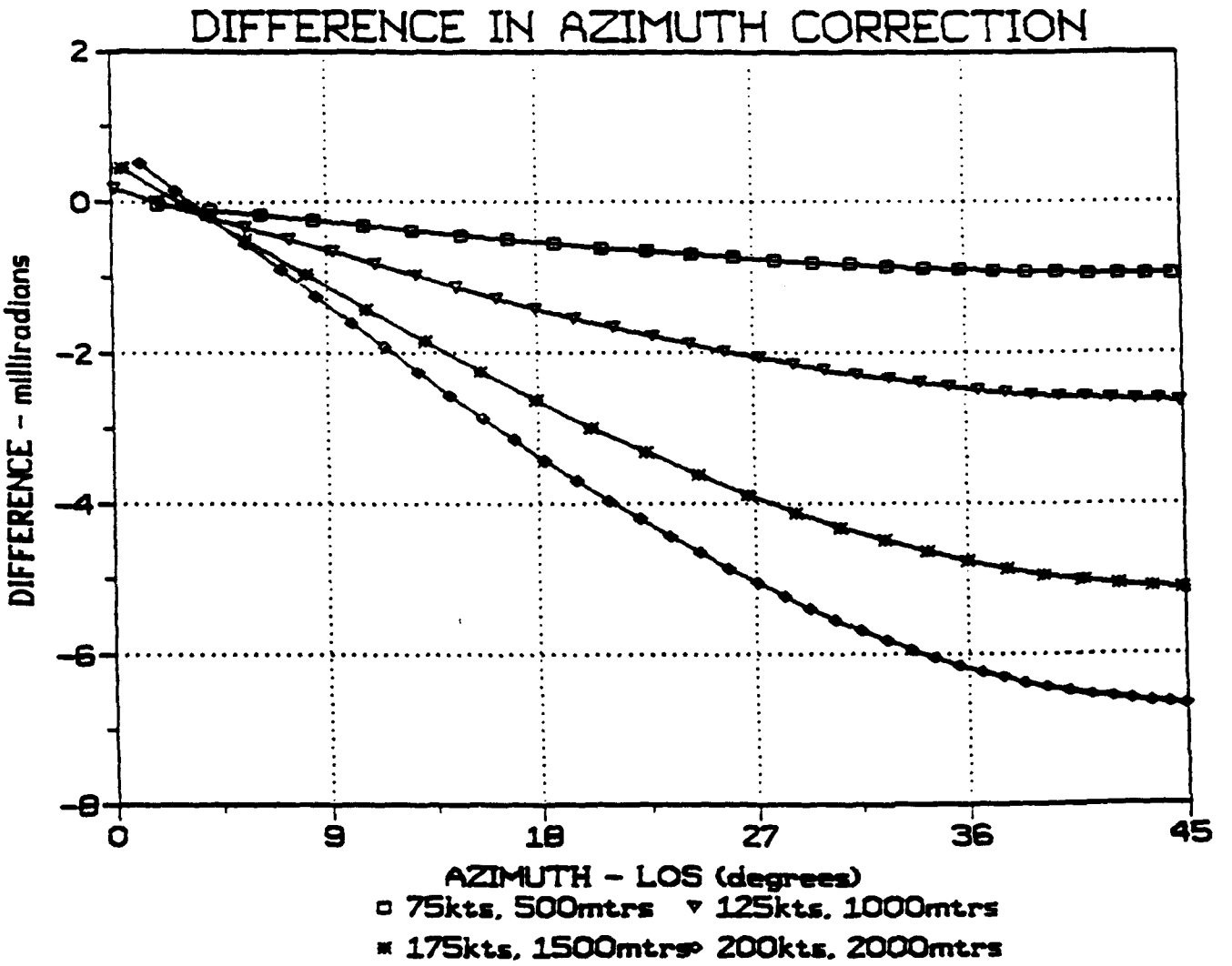


Figure II-3

Azimuth Aiming Error Due to Mounting Rate Sensor
Along Gun Boresight Rather Than Laser Beam Centerline
(Flight starts at 45 degrees azimuth and ends at 0 degrees)

From figure II-3 it can be seen that aiming error when rates are measured along gun boresight versus along laser beam centerline in the azimuth plane is substantial especially at high platform velocities and large azimuth angles.

THEORETICAL TECHNIQUE TO REDUCE AIMING ERROR

If equation II-1 is differentiated with time

$$\dot{AZ}_{gun} = \dot{AZ}_{laser} + \dot{AZ}_{corr} \qquad \text{(II-7)}$$

if $k = V_{muz} \, Cos(EL_{corr})/RSL$

from equation II-2

$$\dot{AZ}_{laser} = kAZ_{corr}$$

therefore

$$\dot{AZ}_{gun} = kAZ_{corr} + \dot{AZ}_{corr}$$

or

$$\dot{AZ}_{corr} = -kAZ_{corr} + \dot{AZ}_{gun} \qquad \text{(II-8)}$$

which has the solution

$$AZ_{corr} = Qe^{-kt} + \dot{AZ}_{gun}/k$$

where t is the sampling period time interval.

As a first try for Q, assume

at $t = 0$, AZ_{corrp} (the previous correction)

therefore

$$Q = AZ_{corrp} - \dot{AZ}_{gun}/k$$

finally

$$AZ_{corr} = AZ_{corrp} e^{-kt} - (e^{-kt} - 1) AZ_{gun}/k \qquad \text{(II-9)}$$

Figure II-4 shows the improvement in azimuth aiming correction when azimuth correction defined by equation II-9 is used.

The error can be further improved if a "correct" correction is used. For example, this can be determined by by locking the laser beam wrt gun during an initial computation period and using this as the initial condition. The initial correction then becomes

$$AZ_{corr} = AZ_{initial} e^{-kt} - (e^{-kt} - 1) AZ_{gun}/k \qquad \text{(II-10)}$$

Subsequently, the previous solution would be used (i.e. equation II-9). The improvement in aiming error is now shown in figure II-5.

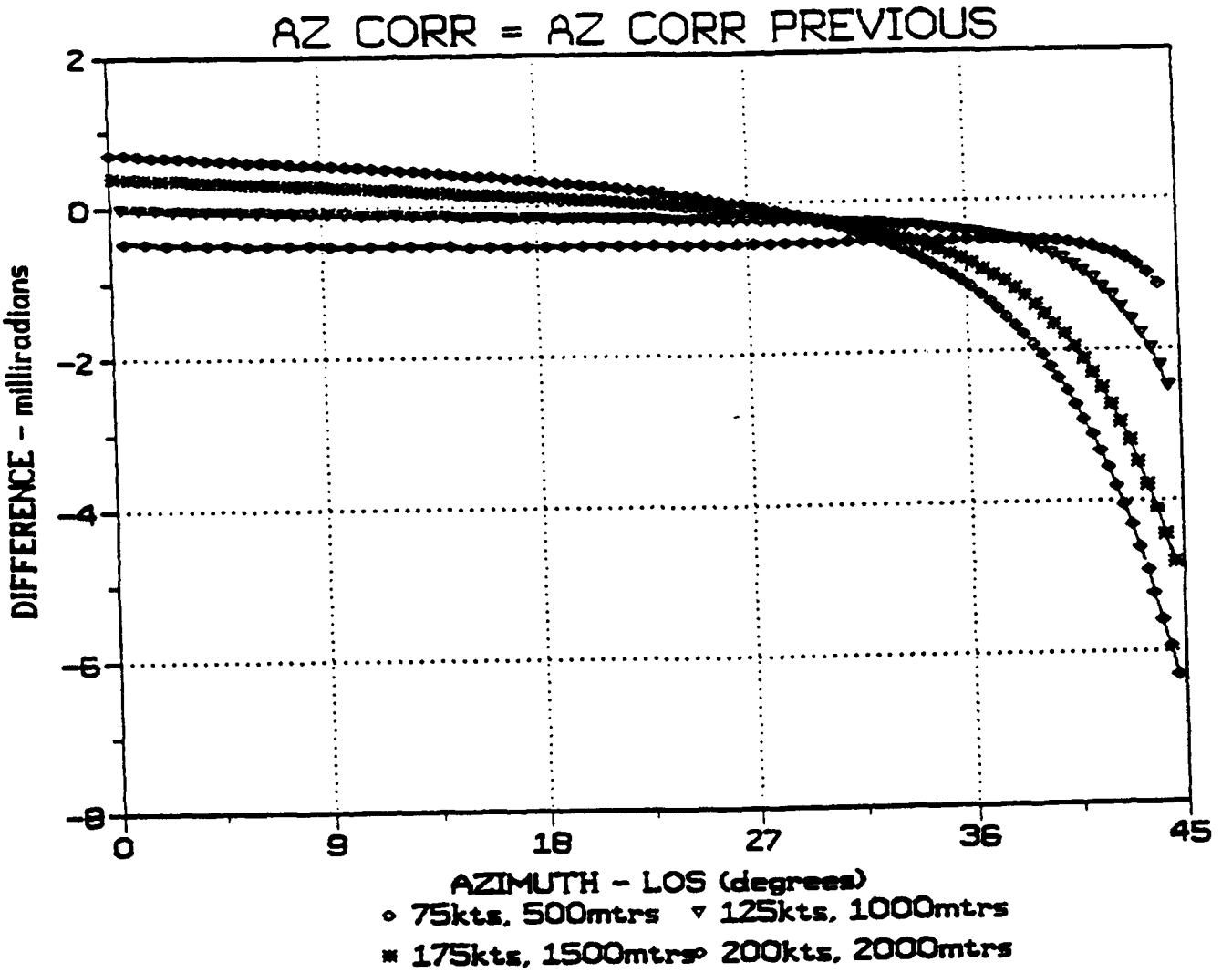


Figure II-4
Improvement in Azimuth Aiming Correction When
Equation II-9 is Used

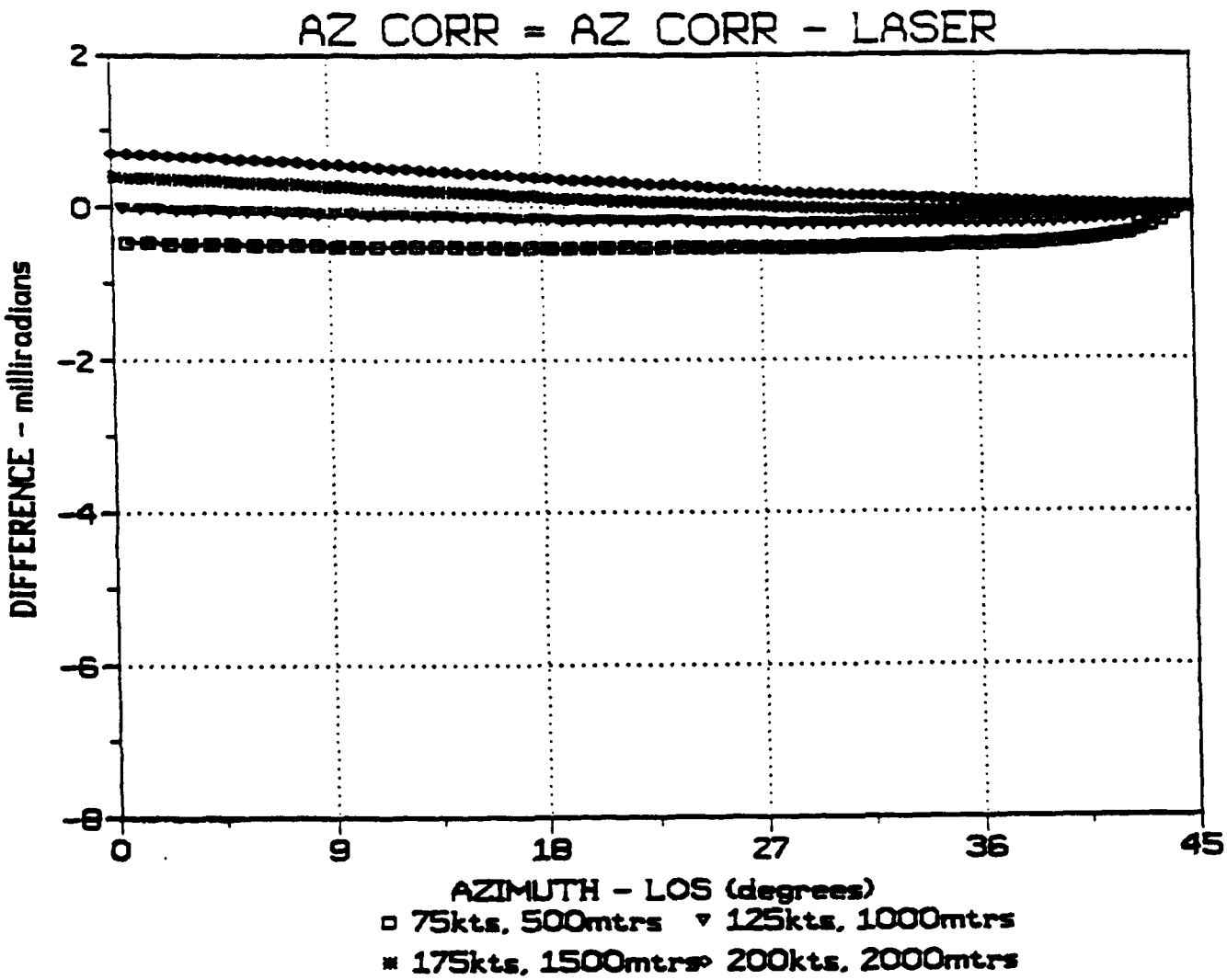


Figure II-5
Improvement in Azimuth Aiming Correction When
Equation II-9 and II-10 are Used

The logic of this correction technique was applied to the BSTING fire control algorithms. First, for the first three seconds, range, azimuth rate and elevation rate were used to compute the azimuth and elevation corrections. The gun and laser beam moved together during this period. Next, one second was used to implement the solution and reaim the beam onto the target. Finally, the gun was allowed to move independent of the laser beam, but still used to keep the beam on the target. The correction algorithm described by equations II-9 and II-10 were used. The value of $AZ_{initial}$ was the value of the aiming correction solution at the end of the calculation period. The following scenario was used:

A/C VELOCITY	RANGE AT CLOSEST APPROACH	ALTITUDE
150 knots	1000 mtrs	250 feet

Figure II-6 shows the gun and laser beam azimuth rates. Incidently, this correction technique was not applied to the elevation rates since the difference in elevation rate between gun and laser are significantly less, though the correction could be applied here as well.

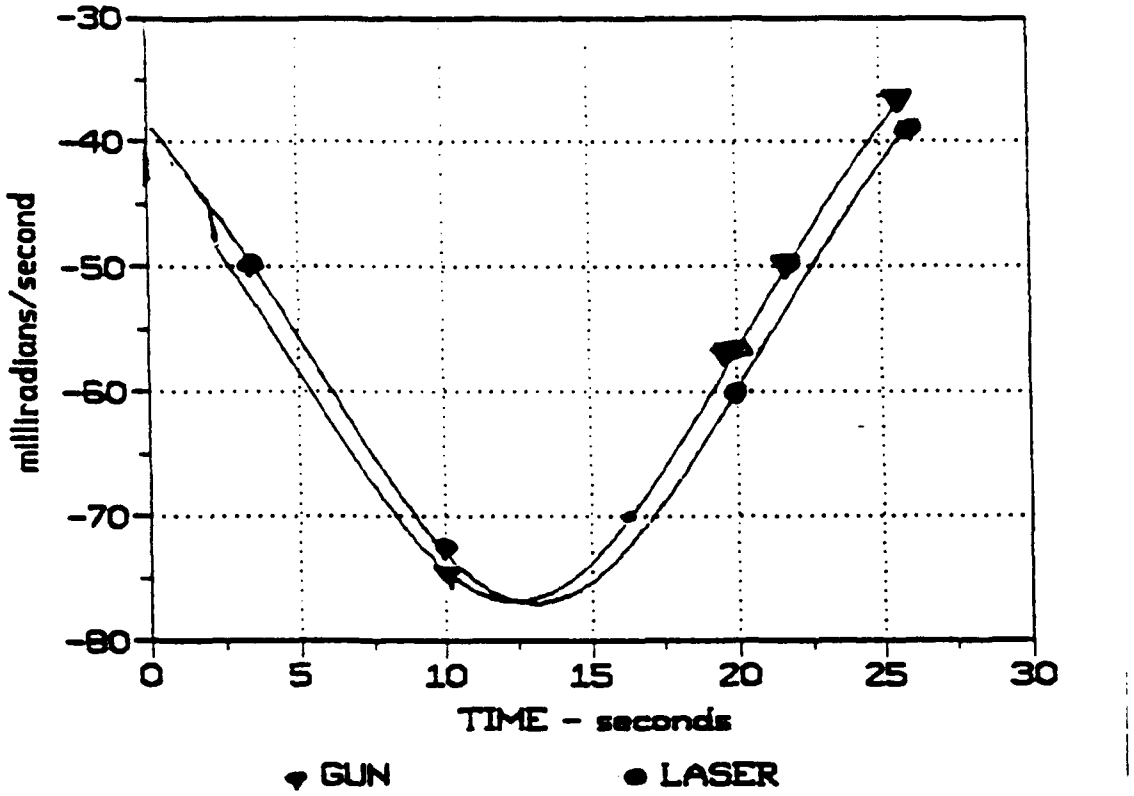
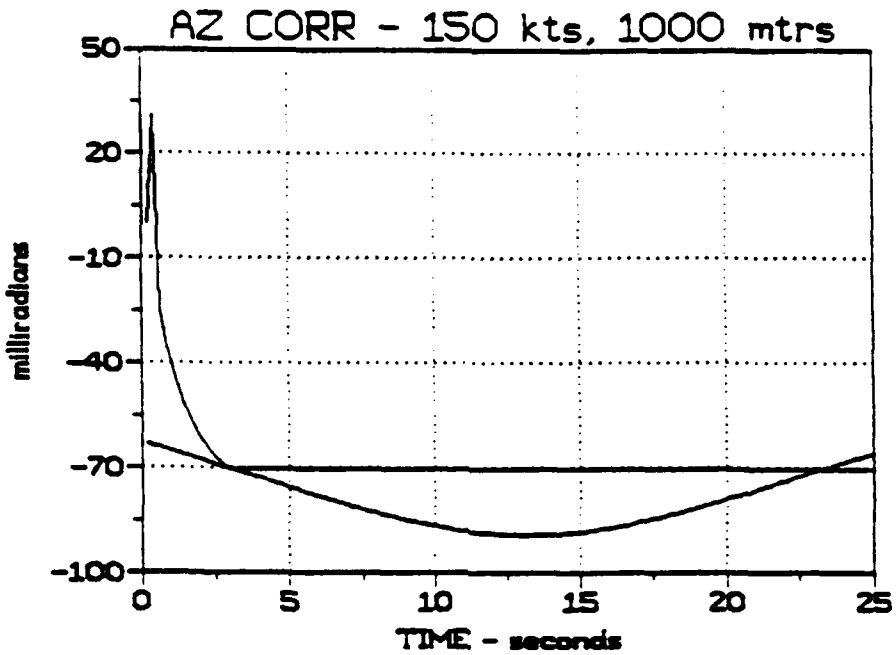


Figure II-6
Azimuth Rates "Measured" Along Gun Boresight and Laser Beam Difference Results Because of the Contantly Changing Aiming Correction

Figure II-7 shows the azimuth and elevation corrections that would be applied with the current BSTING algorithm. Figure II-8 shows the azimuth and elevation corrections that would be available using continuously updated rate measurements.

The analysis shows that at least theoretically the BSTING fire control system could be modified by revising the algorithms so that they could provide continuously updated aiming corrections.



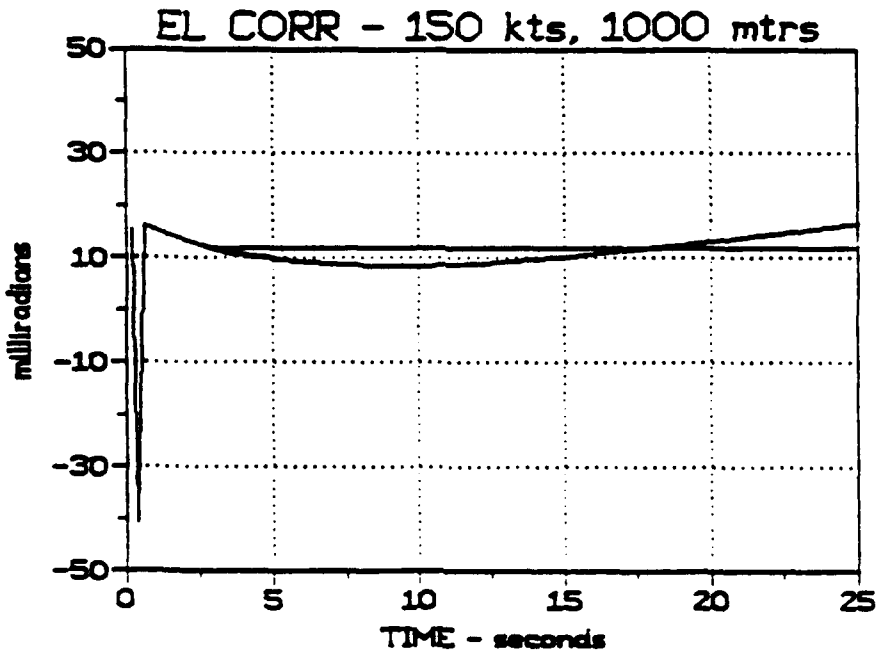
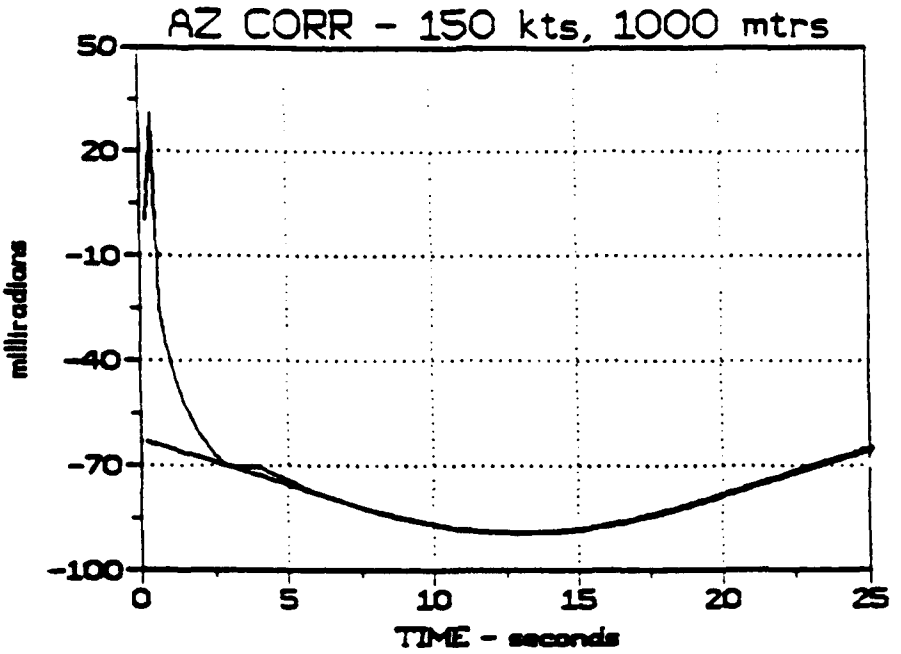


Figure II-7
Azimuth and Elevation Corrections Using Current BSTING Algorithms (Non-noisy Measurements)



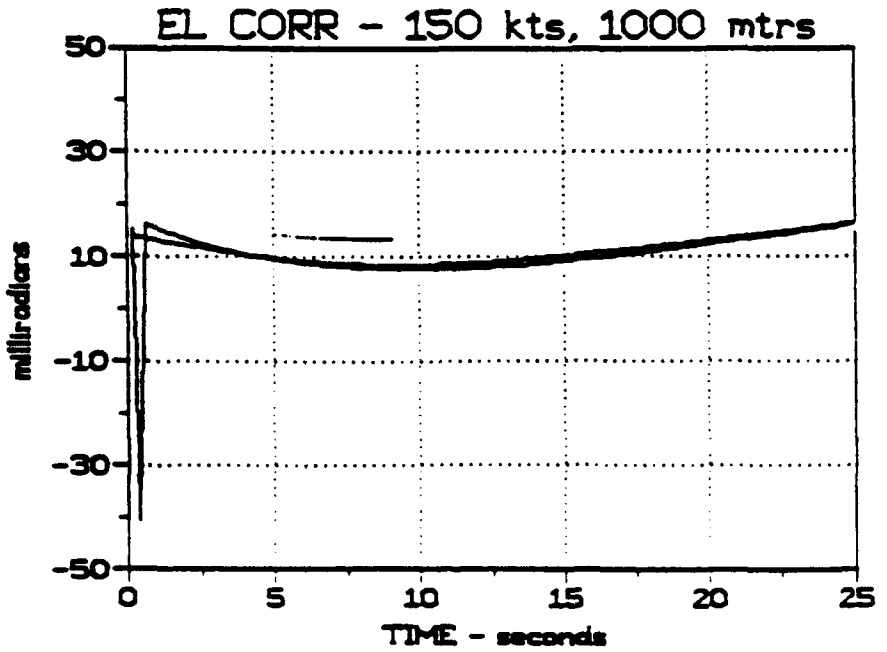


Figure II-8

Azimuth and Elevation Corrections Using Current Modified BSTING Algorithms (Non-noisy Measurements)

COMPUTER PROGRAM: DIFFPR.BAS

PURPOSE: Computes Azimuth Aiming Error Due to Mounting Rate Sensor Along Gun Boresight Rather Than Laser Beam Centerline.

RESULTS: Figure II-3

```
10 OPEN "O", #1, "B:AZDIFF.DAT"
20 PI = 3.1416
30 INPUT "Aircraft Velocity  (knots) =?", VAC
40 INPUT "Range at Closest Approach (meter) =?", RMIN
45 RMIN = RMIN*3.28
50 INPUT "Altitude (feet)  = ?", H
60 INPUT "Muzzle Velocity (fps) = ?", VMUZ
70 VA = VAC*88/60/.8684
80 INPUT "Time Increment = ?", DT
90 RBASE = RMIN - VA*T
100 RSL = SQR(RMIN^2 + RBASE^2 + H^2)
110 AZRATE = -VA*RMIN/RSL/RSL
120 AZLOS = RBASE/RSL
130 AZLOS = ATN(AZLOS/SQR(1 - AZLOS^2))
140 AZLOS1 = AZLOS*180/PI
150 AZCORR = RSL*AZRATE/VMUZ
160 IF AZLOS < 0 THEN GOTO 210
170 DIFF = VA*(COS(AZLOS) - COS(AZLOS + AZCORR))/VMUZ
180 PRINT T, RSL, AZLOS1, AZRATE*1000, AZCORR*1000, DIFF*1000
185 PRINT #1, AZLOS1, DIFF*1000
190 T = T + DT
200 GOTO 90
210 END
```

COMPUTER PROGRAM: ACTCORR.BAS

PURPOSE: Computes the Theoretical Continous Azimuth Aiming Solution Using Either Equations II-9 and II-10. (Depending on the Initial Correction Used)

RESULTS: II-4 and II-5.

```
010 OPEN "I", #1, "B:AZDIFF75.DAT"
20 OPEN "O", #2, "B:AZDIFF75.OUT"
30 PI = 3.1416
40 INPUT "Muzzle Velocity (fps) = ?", VMUZ
50 INPUT "Elevation Correction (degrees) = ?", ELCORR
60 ELCORR = ELCORR*PI/180
65 INPUT #1, T, RSL, AZLOS, AZRATE, AZDOTGUN, AZCORR , DIFF
70 AZCORR1 = AZCORR
80 INPUT #1, T, RSL, AZLOS, AZRATE, AZDOTGUN, AZCORR , DIFF
90 K = VMUZ*COS(ELCORR)/RSL
95 AZCORR1 = AZCORR1*EXP(-.2*K) - (EXP(-.2*K) - 1)*AZDOTGUN/K
122 DIFF1 = AZCORR1 - AZCORR
130 PRINT AZLOS, AZCORR*1000, AZCORR1*1000, DIFF*1000, DIFF1*1000
134 PRINT #2, AZLOS, DIFF1*1000
140 GOTO 80
```

COMPUTER PROGRAM: AZDATAPR.BAS

PURPOSE: Computes the Input Data Used for the Theoretical Continous Aiming Solution Calculated by Equations II-9 and II-10.

RESULTS: Inputs for Programs Used for Figures II-4 and II-5.

```
010 OPEN "O", #1, "B:AZDIFF75.DAT"
20 PI = 3.1416
30 INPUT "Aircraft Velocity  (knots) =?", VAC
40 INPUT "Range at Closest Approach (meter) =?", RMIN
50 RMIN = RMIN*3.28
60 INPUT "Altitude (feet)  = ?", H
70 INPUT "Muzzle Velocity (fps) = ?", VMUZ
80 VA = VAC*88/60/.8684
90 INPUT "Time Increment = ?", DT
100 RBASE = RMIN - VA*T
110 RSL = SQR(RMIN^2 + RBASE^2 + H^2)
120 AZRATE = -VA*RMIN/RSL/RSL
130 AZLOS = RBASE/RSL
140 AZLOS = ATN(AZLOS/SQR(1 - AZLOS^2))
150 AZLOS1 = AZLOS*180/PI
160 AZCORR = RSL*AZRATE/VMUZ
170 IF AZLOS < 0 THEN GOTO 240
180 AZDOTGUN = -VA*COS(AZLOS + AZCORR)/RSL
190 DIFF = VA*(COS(AZLOS) - COS(AZLOS + AZCORR))/VMUZ
200 PRINT T, RSL, AZLOS1, AZRATE*1000, AZDOTGUN*1000, AZCORR*1000, DIF!
210 PRINT #1, T, RSL, AZLOS1, AZRATE, AZDOTGUN, AZCORR, DIFF
220 T = T + DT
230 GOTO 100
240 END
```

CHAPTER III

SENSITIVITY ANALYSIS

BSTING System

A brief investigation was done to determine the sensitivity of aiming corrections to the accuracy of the sensor measurements. Because the correction in the azimuth plane due to movement of the gun platform with respect to the target is most critical, and the largest source of error, this parameter was used in determining sensitivities.

Since the algorithms used in the BSTING System represents probably the simplest technique possibility to determine the correction they were used in this investigation. In the BSTING system, the azimuth aiming correction is

$$\sin(AZ_{corr}) = -RSL \times \dot{AZ}/V_{muz} \qquad \text{(III-1)}$$

where: RSL = Slant Range to Target

$\dot{AZ}$ = Azimuth Rate Measured by Rate Sensor

V_{muz} = Muzzle Velocity

During measurements the gun is locked to the laser beam so that

$$\dot{AZ}_{laser} = \dot{AZ}_{gun} = \dot{AZ}$$

Therefore the azimuth correction is sensitive to three parameters - slant range, azimuth rate and muzzle velocity. This dependence can be expressed by the relationship:

$$AZ_{corr} = \frac{\partial Corr}{\partial RSL}\Delta RSL + \frac{\partial Corr}{\partial \dot{AZ}}\Delta \dot{AZ} + \frac{\partial Corr}{\partial V_{muz}}\Delta V_{muz} \qquad \text{(III-2)}$$

Using equation III-1, the sensitivity coefficients can be expressed as:

PARAMETER	DESCRIBING EQUATION
$\frac{\partial Corr}{\partial RSL}$	$TERM \times \dot{AZ}/V_{muz}$
$\frac{\partial Corr}{\partial AZ}$	$TERM \times RSL/V_{muz}$
$\frac{\partial Corr}{\partial V_{muz}}$	$TERM \times RSL \times \dot{AZ}/V_{muz}^{2}$

$$TERM = \frac{1}{\sqrt{1 - (RSL \times AZ/V_{muz})^{2}}}$$

The above sensitivities were computed for several scenarios. In each case the helicopter flew past the target on a straight and level trajectory. The scenarios used are shown in table III-1.

TABLE III-1

SCENARIOS USED IN SENSITIVITY ANALYSIS

SCENARIO NO.	VELOCITY (knots)	RANGE AT CLOSEST APPROACH (meters)	ALTITUDE (feet)
1	75	500	250
2	150	500	250
3	175	1000	250
4	175	1500	250
5	175	2000	250
6	175	2500	250

The resulting sensititivies are shown in figures III-1 through III-3.

In order to determine the error contribution on each term in equation III-2 it was necessary to estimate the true measurement accuracy of each of the three sensors used in computing the azimuth correction. The true accuracy includes not only the inherent accuracy of the sensor itself but also the actual accuracy that can be achieved when the sensor is aimed by the gunner on the moving platform. This "total" sensor accuracy was determined using representative data from the BSTING flight tests conducted at the Naval Weapons Center during August and September of 1990.

A 10 meter/second (32.8 feet/second) is assumed for muzzle velocity variation as suggested by Reference III-1. According to the reference, the main cause of variation in muzzle velocity is due to variations in propellant temperature. Thus this error could be reduced by including a sensor that measures the temperature in the ammunition container. Referring to figure III-2, a 10 m/s error in muzzle velocity would result in no more than about a 1-milliradian error in azimuth correction.

Review of flight test data showed that when the gunner had the laser beam aimed at the target, the range error is no more than 10 feet. Figure III-4 that shows the results from Pass No. 6 on August 6, 1990, which is typical. This is unfiltered data and thus filtered data that is actually usedin calculating corrections is even better. Combining this with the sensitivities in figure III-1, the maximum error in azimuth aiming correction due to range measurements is probably no more than 0.5 milliradian.

After subtracting the muzzle velocity error (1 milliradian) and range measurement (0.5 milliradian) from the overall 4 milliradian criteria, 2.5 milliradians can be alloted to rate azimuth rate sensor measurements. Using a sensitivity of 3 milliradians/mr-per-second (see figure III-3), a rate sensor accuracy of about 0.8 milliradians-per-second would be required.

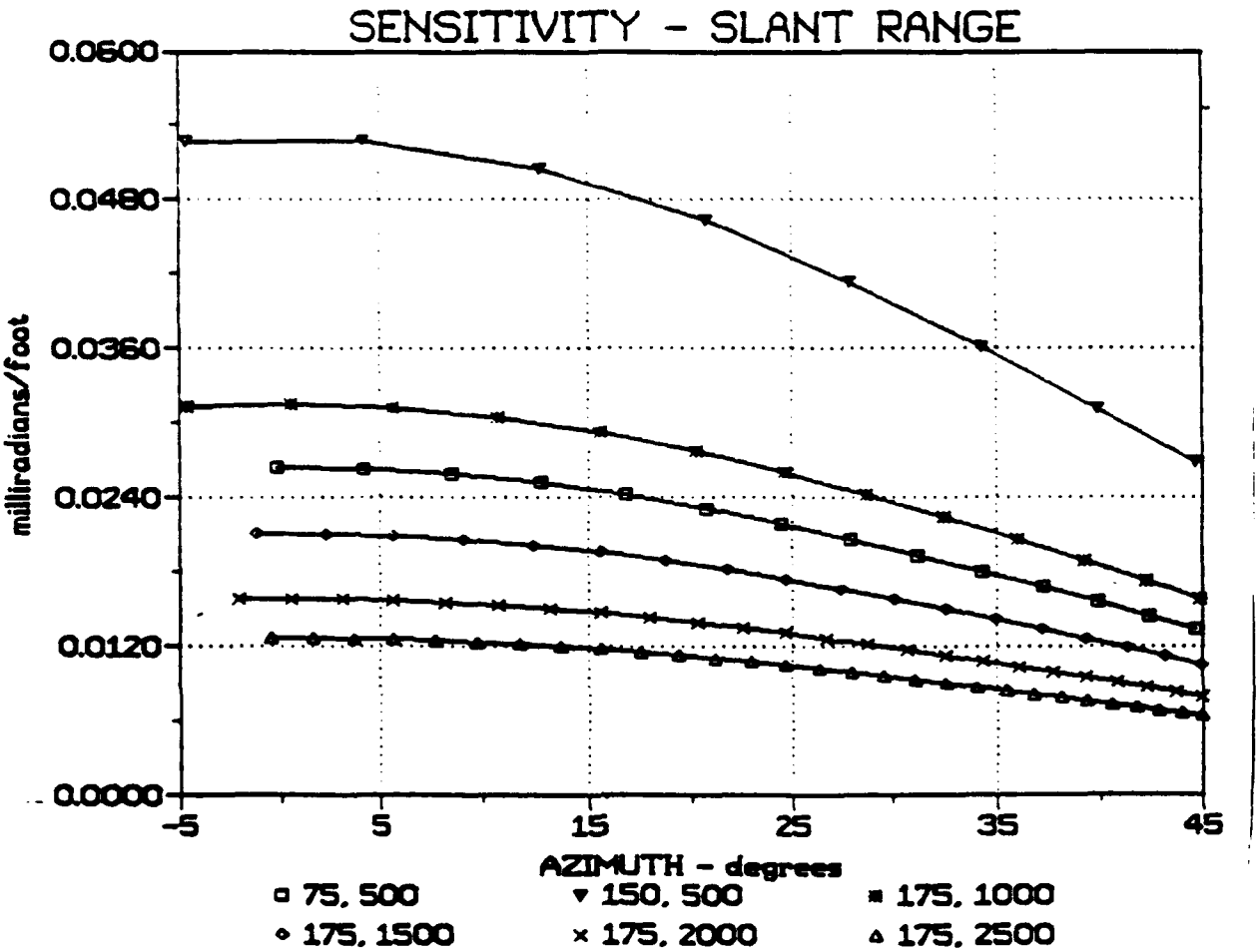


Figure III-1

Sensitivity of the Correction in Azimuth Due to Variations in Slant Range Measurements. The Two Numbers in the Legend are Platform Velocity in knots and Slant Range in meters. (Flight starts at 45 degrees azimuth and ends at 0-5 degrees)

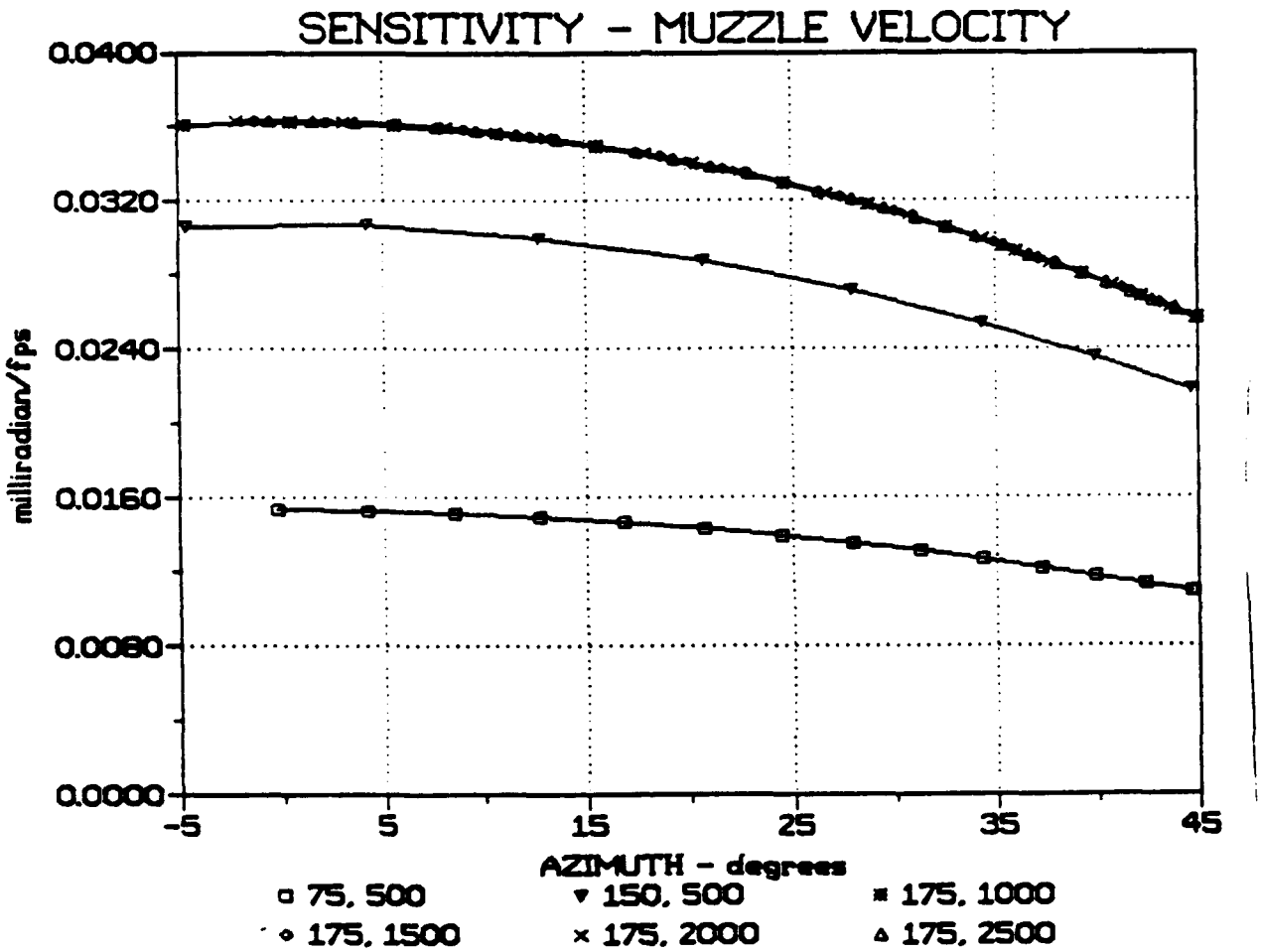


Figure III-2

Sensitivity of the Correction in Azimuth Due to Variations in Muzzle Velocity Used in Computations. The Two Numbers in the Legend are Platform Velocity in knots and Slant Range in meters. (Flight starts at 45 degrees azimuth and ends at 0-5 degrees)

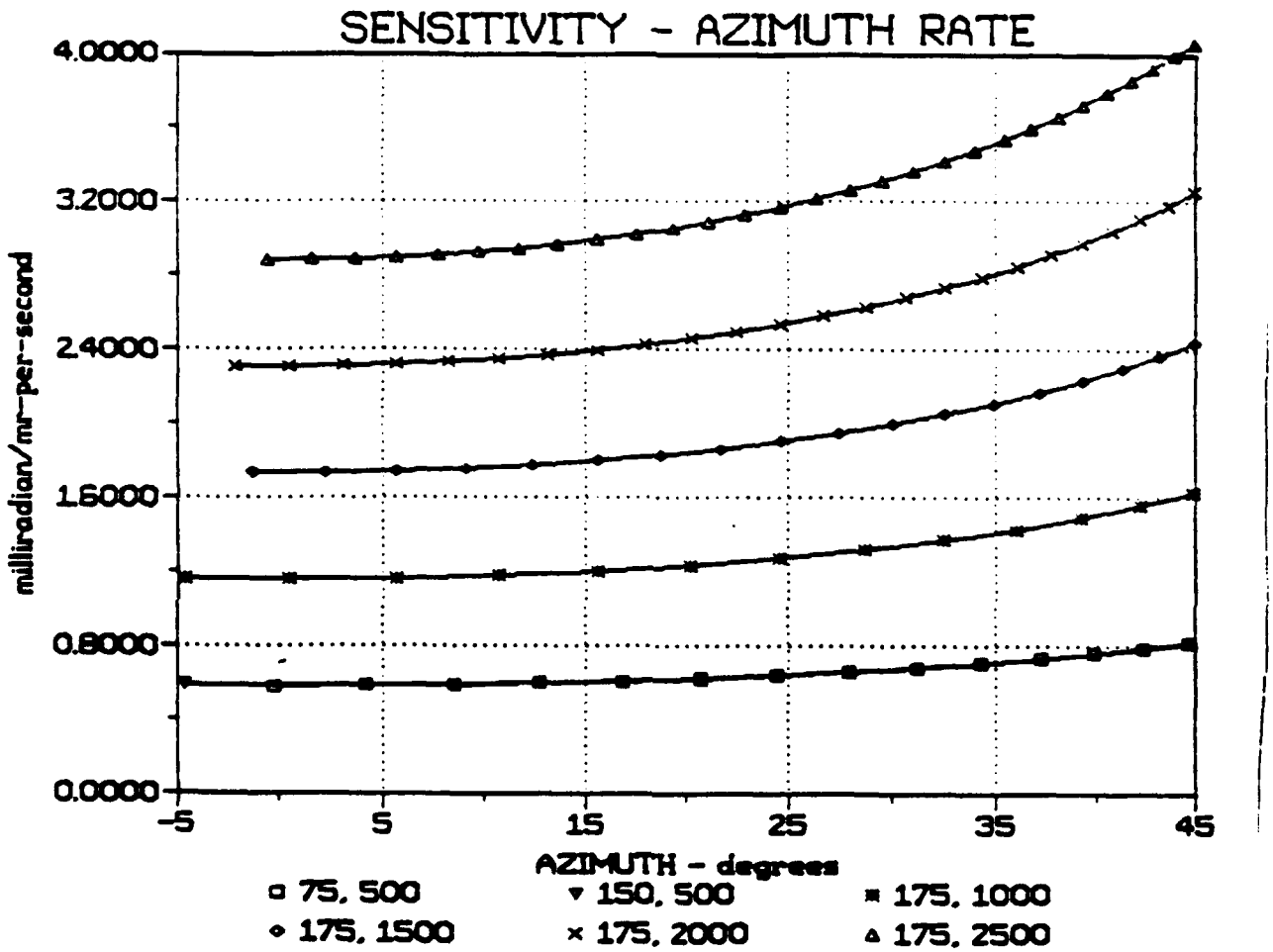


Figure III-3

Sensitivity of the Correction in Azimuth Due to Variations in Azimuth Rate Measurements. The Two Numbers in the Legend are Platform Velocity in knots and Slant Range in meters. (Flight starts at 45 degrees azimuth and ends at 0-5 degrees)

Greater accuracy is required as range increases. For example, in the BSTING flight tests, the range was limited to usually 1000 meters. therefore looking at figure III-3, an accuracy of about 2-milliradians-per-second was acceptable.

Figure III-5 shows filtered azimuth rate used in the azimuth correction versus that actually required for a representative case, again for Pass 6 on August 7, 1990. This shows that the gunner is able to aim the gun/laser so that a 2-milliradians-per-second rate sensing accuracy is possible.

While longer ranges were not used during the flight tests, a qualitive statement can be made. Since at longer ranges result in reduced azimuth rates, it is easier to keep the laser spot on target so rate sensing errors should be substantially less. Therefore this should, at least partially, compensate for the requirements that rate sensing accuracy must be better at longer ranges.

REFERENCES

III-1. Fire Control Analysis of Selected Crew Served Weapons, Thomas B. Knapp and K.P. Pfleger, U.S. Army Armament Research, Development and Engineering Center, TR ARFSD-TR-90032, March 1991.

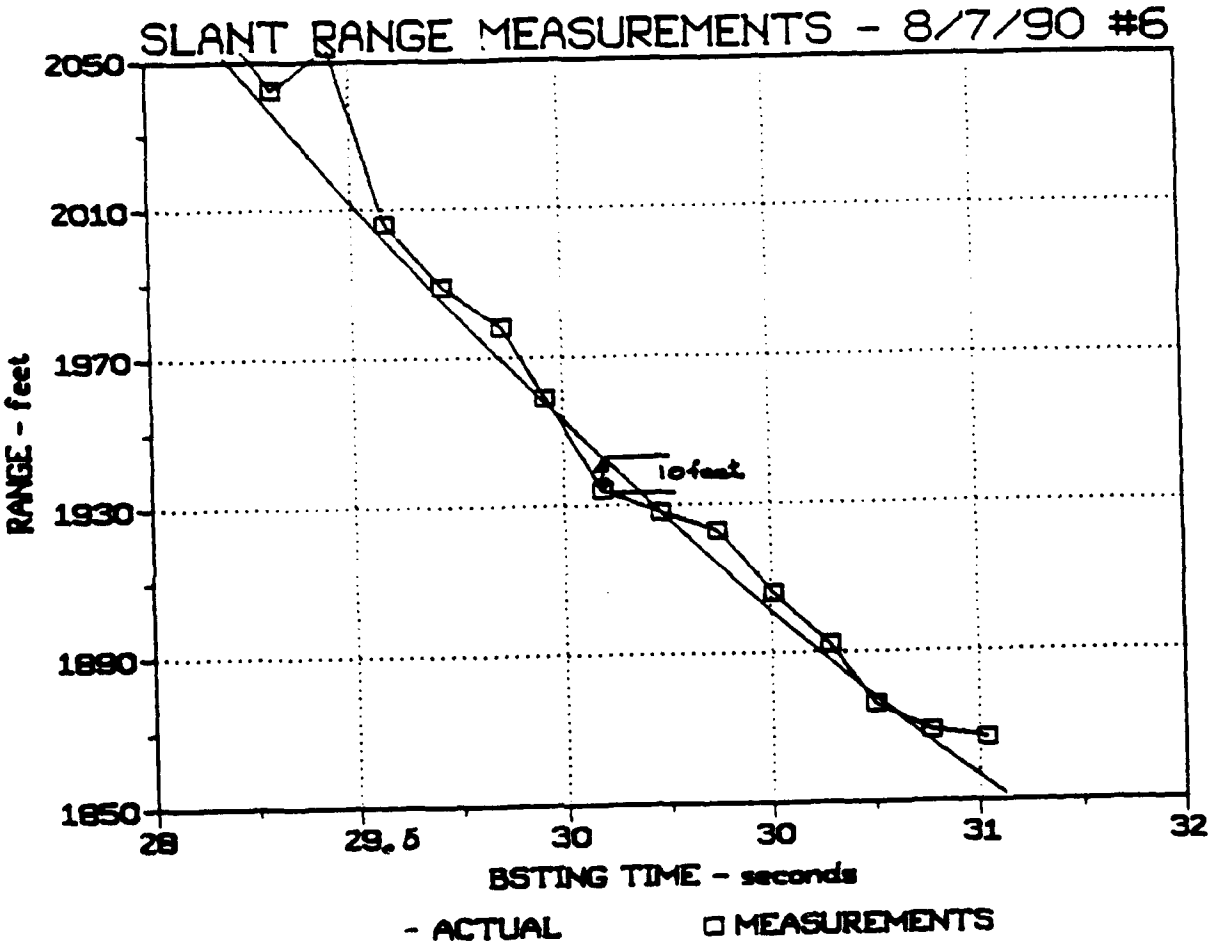


Figure III-4

Comparison of Actual Slant Range Distance with
Unfiltered Range Measurement Data
(Pass Number 6 - 8/7/90)

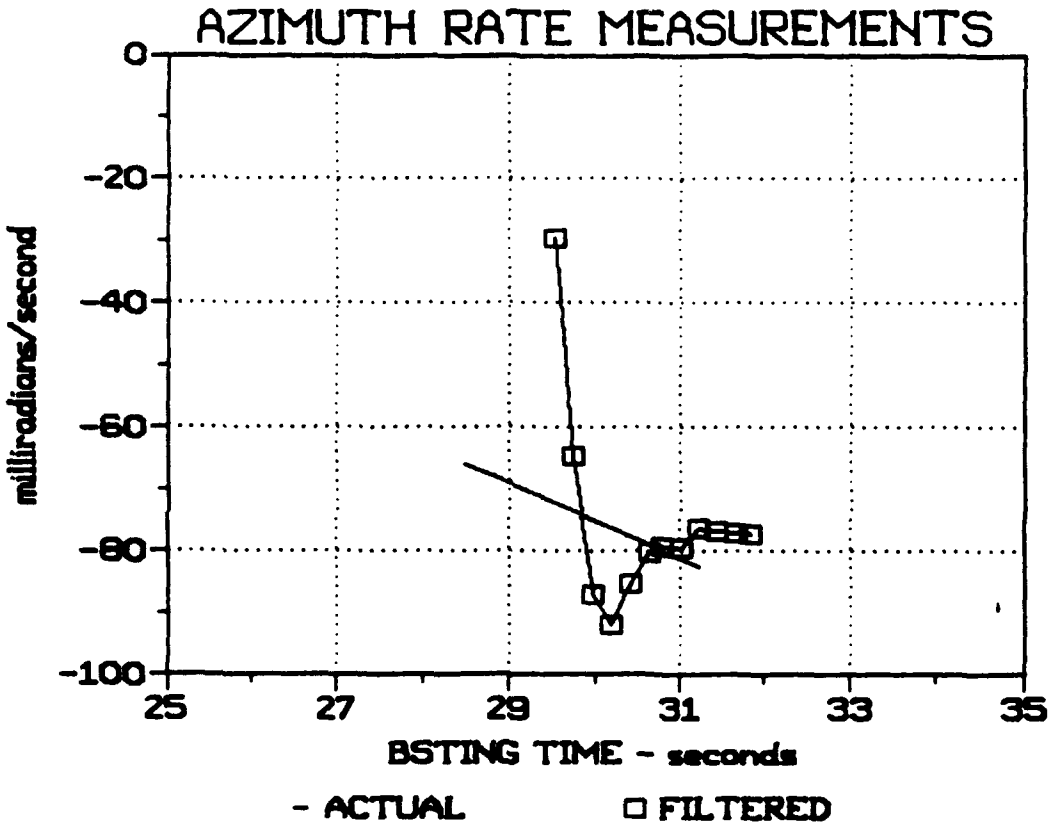


Figure III-5

Comparison of Actual Azimuth Rate with Filtered Azimuth Rate Sensor Data)
(Pass Number 6 - 8/7/90)

COMPUTER PROGRAM: SENSPR.BAS

PURPOSE: Computes the sensititivies of the azimuth correction to variations in (1) muzzle velocity, (2) slant range measurements and (3) azimuth rate measurements.

RESULTS: III-1, III-2 and III-3

```
 10 OPEN "O", #1, "B:CORRRSL2.DAT"
20 OPEN "O", #2, "B:CORRRTE2.DAT"
30 OPEN "O", #3, "B:CORRMUZ2.DAT"
40 PI = 3.1416
50 G = 32.17
60 INPUT "Aircraft Velocity (knots) = ?", VAC
70 VAC = VAC*88/60/.8684
80 INPUT "Range at Closest Approach (meters) = ?", RMIN
90 RMIN = RMIN*3.28
100 INPUT "Altitude (feet) = ?", H
110 RBASE = RMIN - VAC*T
120 RSL = SQR(RMIN^2 + RBASE^2 + H^2)
130 AZ = RBASE/RSL
140 AZ = ATN(AZ/SQR(1 - AZ^2))
150 AZ1 = AZ*180/PI
160 VMUZ = 2860
170 AZRATE =  VAC*RMIN/RSL/RSL
180 COEF = RSL*AZRATE/VMUZ
190 COEF = 1/SQR(1 - COEF^2)
200 DCDRSL = COEF*AZRATE/VMUZ
210 DCDRATE = COEF*RSL/VMUZ
220 DCDVMUZ = COEF*RSL*AZRATE/VMUZ^2
230 PRINT T, AZ1, DCDRSL*1000, DCDRATE, DCDVMUZ*1000
240 PRINT #1, AZ1, DCDRSL*1000
250 PRINT #2, AZ1, DCDRATE
260 PRINT #3, AZ1, DCDVMUZ*1000
270 IF AZ1 < 0 THEN GOTO 400
280 T = T +  1
281 GOTO 110
400 END
```

CHAPTER IV

NOISE SIMULATION AND FILTER OPTIMIZATION

Initial Investigation

It was shown in Chapter II that it was possible to use the BSTING concept for obtaining continuously updated aiming corrections, at least when theoretical "unnoisy" range and azimuth/elevation sensor measurements are used. Next, it is necessary to see if this same conclusion can be made in the actual BSTING environment with noisy sensor measurements.

The technique used to simulate the real situation was to superimpose noise on the theoretical range and azimuth/elevation rates. The noise to be superimposed was determined from actual data obtained during the BSTING flight test program. Again Pass No. 6 on August 7, 1990 was used. The noise was obtained using the following steps.

a. A tracking period was chosen from the flight test records where it was obvious that the gunner was doing a good job of keeping the laser spot on the target.

b. For this period chosen, mean values of range and azimuth/elevation rates were determined using a second-order curve fit to the noisy flight test data.

c. These mean values were then subtracted from the noisy data to obtain the "generic" noise required.

The resulting generic noise is shown in figures IV-1 through IV-3.

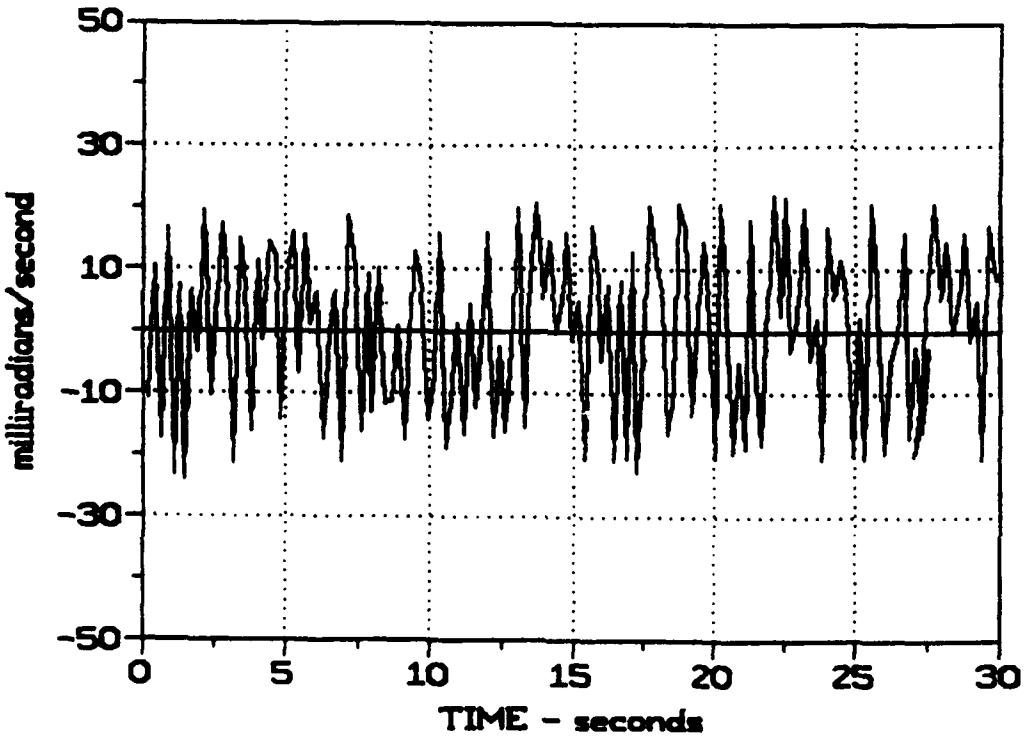


Figure IV-1

Azimuth Rate Generic Noise from Actual Flight Test Data.

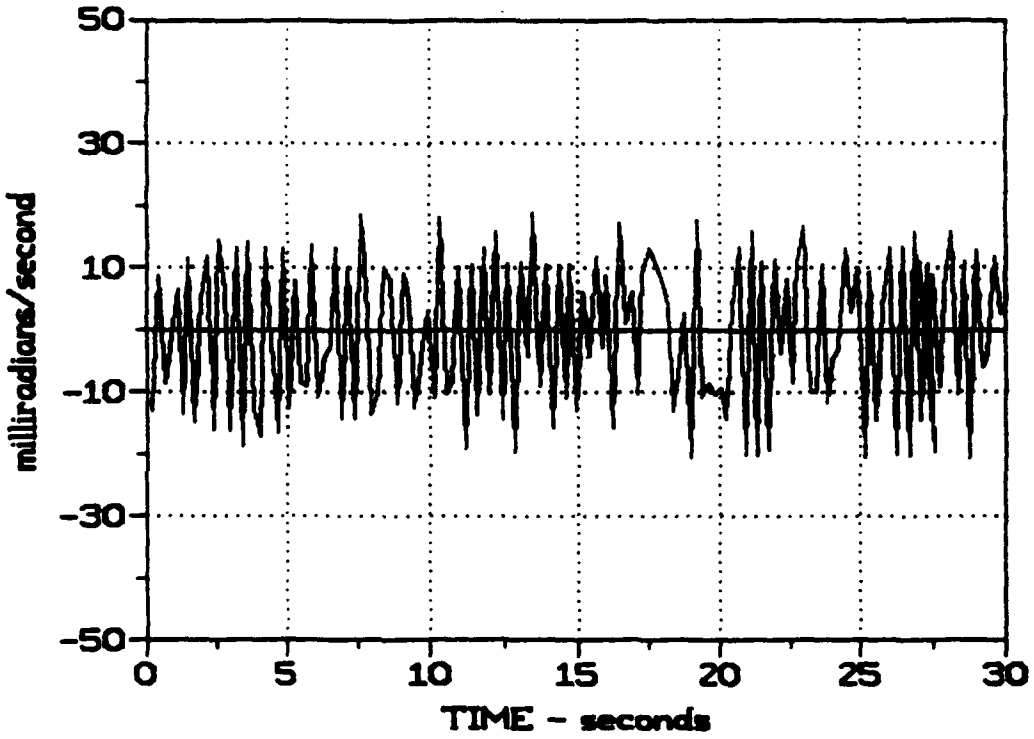


Figure IV-2

Elevation Rate Generic Noise from Actual Flight Test Data.

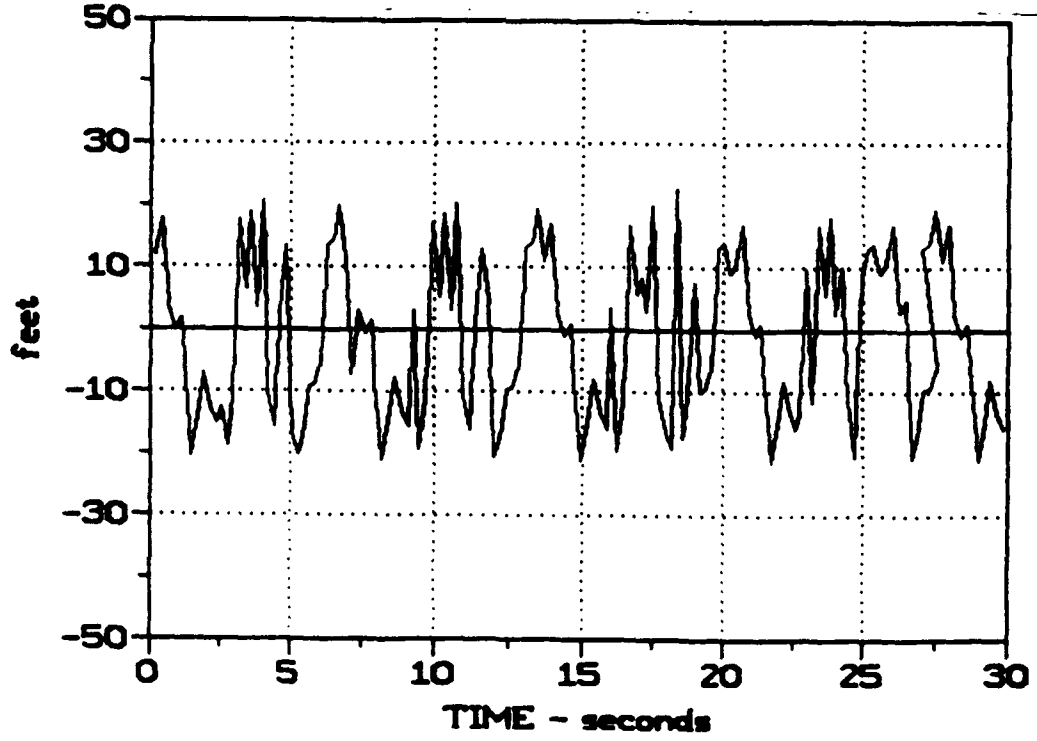


Figure IV-3

Slant Range Generic Noise from Actual Flight Test Data.

A brief analysis was done to determine if the azimuth rate filtering could be improved to give better results than the current filter algorithms now used in the current BSTING system. Incidently, some optimization had been done when the original BSTING algorithms were developed. Since the azimuth correction is the one most sensitive to noise, it was the only one considered here. Further, optimization of either range or elevation rate filters would not significantly affect results. (Note: later the results of using an adaptive filter that constantly provides an optimum filter will be presented.)

The filter algorithm in the azimuth plane is:

```
AZDHAT = TERM2*(1 - 2.0*RHATD*DTA/RHAT)*AZDHAT    (IV-1)
         + KPA*GXHAT
```

Where:

AZDHAT is the filtered azimuth rate

RHAT is the filtered slant range rate

RHATD is the filtered slant range rate

GXHAT is the latest azimuth rate measurement

DTA is time interval between data samples

TERM2 = 0.8 in the current BSTING algorithm

KPA = 0.2 in the current BSTING algorithm

The filter for the slant range measures is:

```
RHAT = C1*RHAT + C2*RHATD + KPR*SENRNG                    (IV-2)

RHATD = RHAT*(TERM3 - KVR) + C3*RHATD + KVR*SENRNG

TERM3 = (ELDHAT^2 + AZDHAT^2)*DTR
```

Where: (in addition to terms previously defined)

SENRNG is the latest range measurement

ELDHAT is the filtered elevation rate

C1 = 0.9

C2 = C1*DTR

DTR is the time interval between data samples

Only one scenario was used in this evaluation and it used the following parameters, again for a straight and level flyby from an azimuth of 45 to -45 degrees:

PLATFORM VELOCITY	-	175 knots
MINIMUM SLANT RANGE	-	1000 meters
ALTITUDE ABOVE TARGET	-	250 feet

Figure IV-4 shows the noisy azimuth rate data when the generic azimuth noise was added to the theoretical azimuth rate. This data would be the raw azimuth rate measurements that would be subjected to different filter parameters in the subsequent optimization investigation.

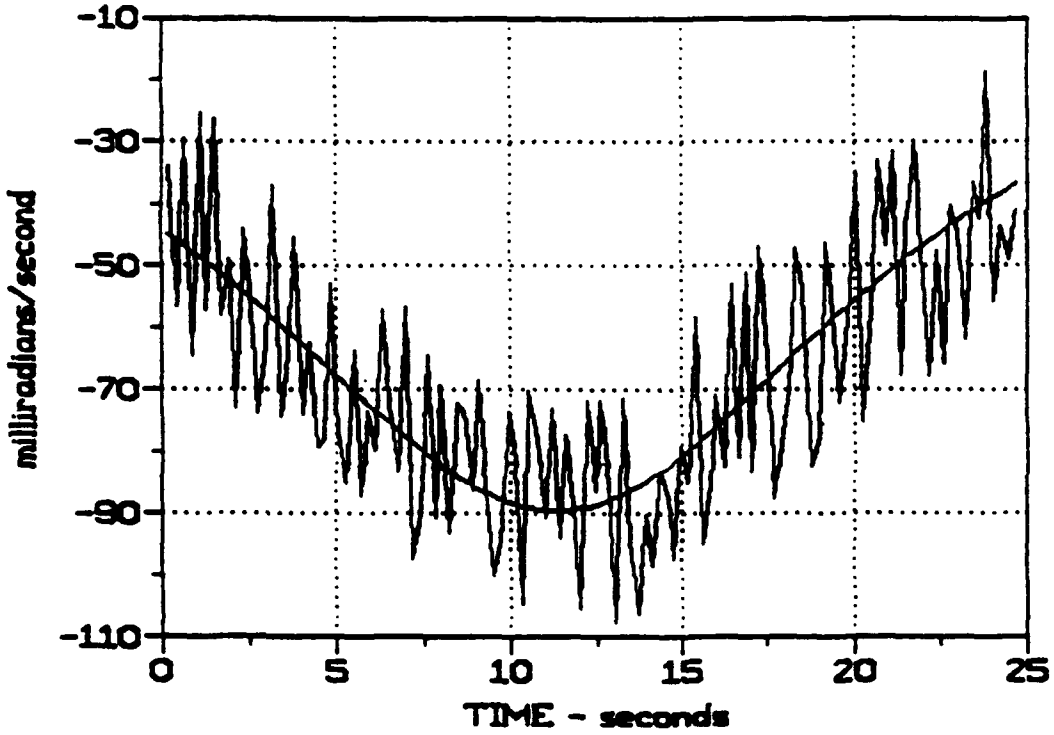


Figure IV-4

"Raw" Azimuth Rate Measurements

In the current BSTING system, the values of KPA and TERM2 are 0.2 and 0.8 respectively. The filtered azimuth rate using these initial coefficient are shown in figure IV-5.

The first improvement consisted of changing the values of KPA and TERM2 to 0.1 and 0.9 respectively. This has the effect of placing more dependence on the "old" correction rather than the "latest" measurement value. This change in coefficient also results in a more sluggish response. Another change made was to

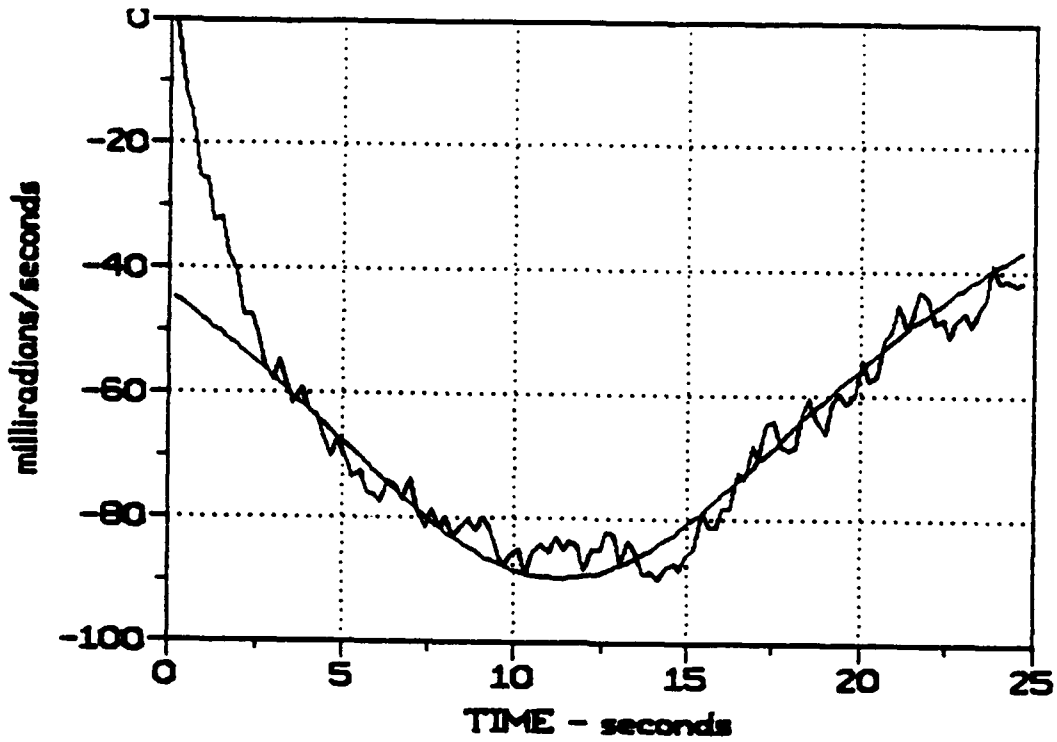


Figure IV-5

Filtered Results - KPA = 2.0, TERM2 = 0.8, GXHAT(0) = 0

change the initial value for the azimuth rate. In the current algorithm, the initial value was assumed to be zero (GXHAT = 0). This was changed so that initial value was taken to be equal to the first azimuth rate measurement. The results of using these new coefficients and initial value for GXHAT are shown in figure IV-6. These changes resulted in a reduced difference between filtered and actual azimuth rates. Also the change in initial condition lead to a reduced initial difference and the filtered data began to oscillate about the actual azimuth rate faster.

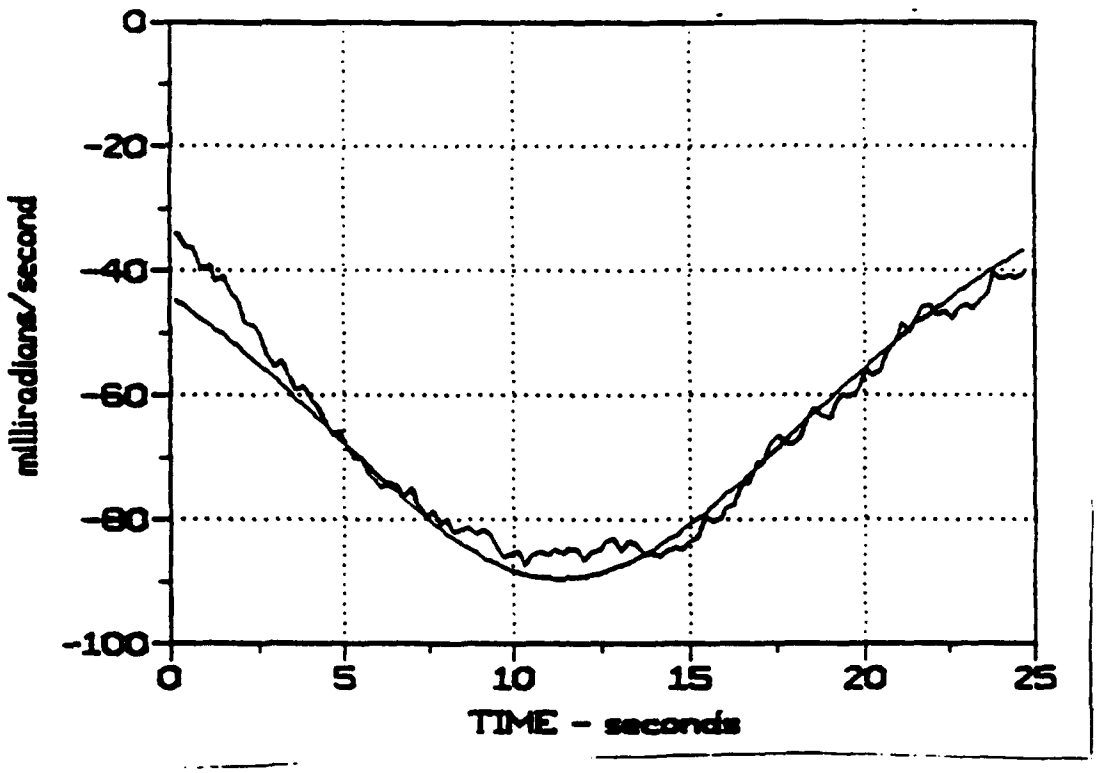


Figure IV-6

Filtered Results - KPA = 1.0, TERM2 = 0.9,
GXHAT(0) = 1st Measurement

Since these changes improved system response, the coefficients were further changed to 0.05 and 0.95. The results of these changes are shown in figure IV-7. Again the initial value was taken to be one-half the initial azimuth rate sensor measurement. This showed a further reduction in oscillation about the nominal rates. However, the sluggishness of the response was too great.

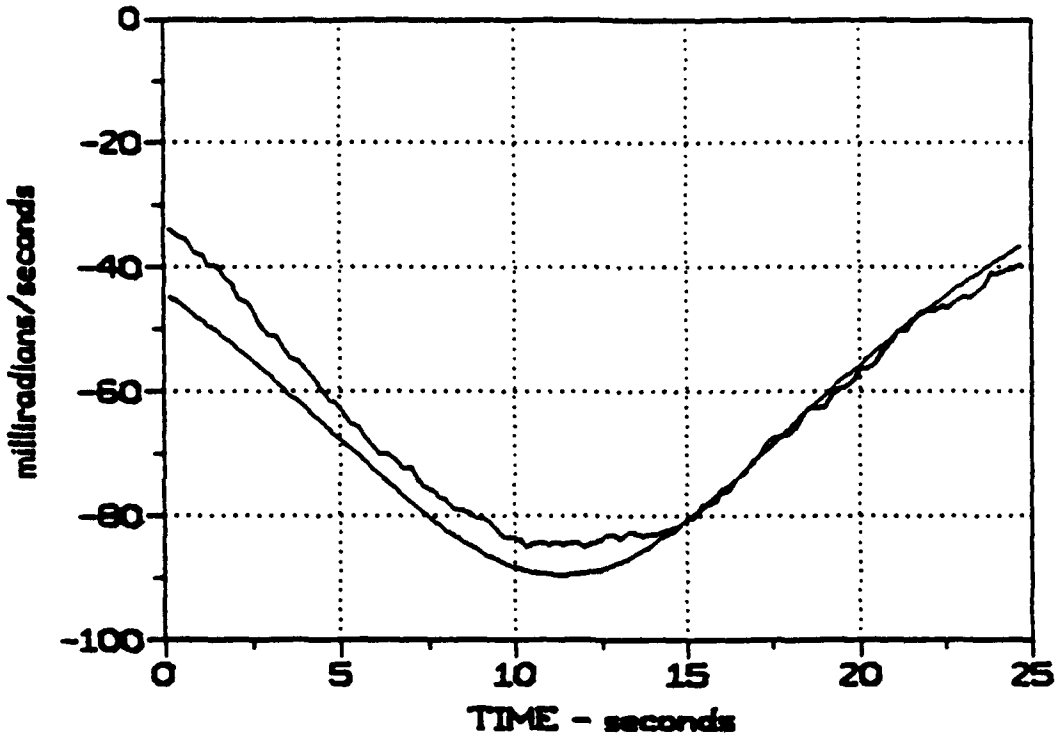


Figure IV-7

Filtered Results - KPA = 0.05, TERM2 = 0.95,
GXHAT(0) = 1st Measurement

The next excursion was to see the effect of increasing the sampling rate. Here the increased sampling rate was simulated by using the same noisy measurement, but the interval between the data point was simply halved. The results of this increased in sampling rate is shown in figure IV-8. The increased sampling rate results in the azimuth rate homing in on the nominal value faster (i.e. about 2.5 versus 5 seconds).

The results of this preliminary showed that there could be benefit in increasing the sampling rate.

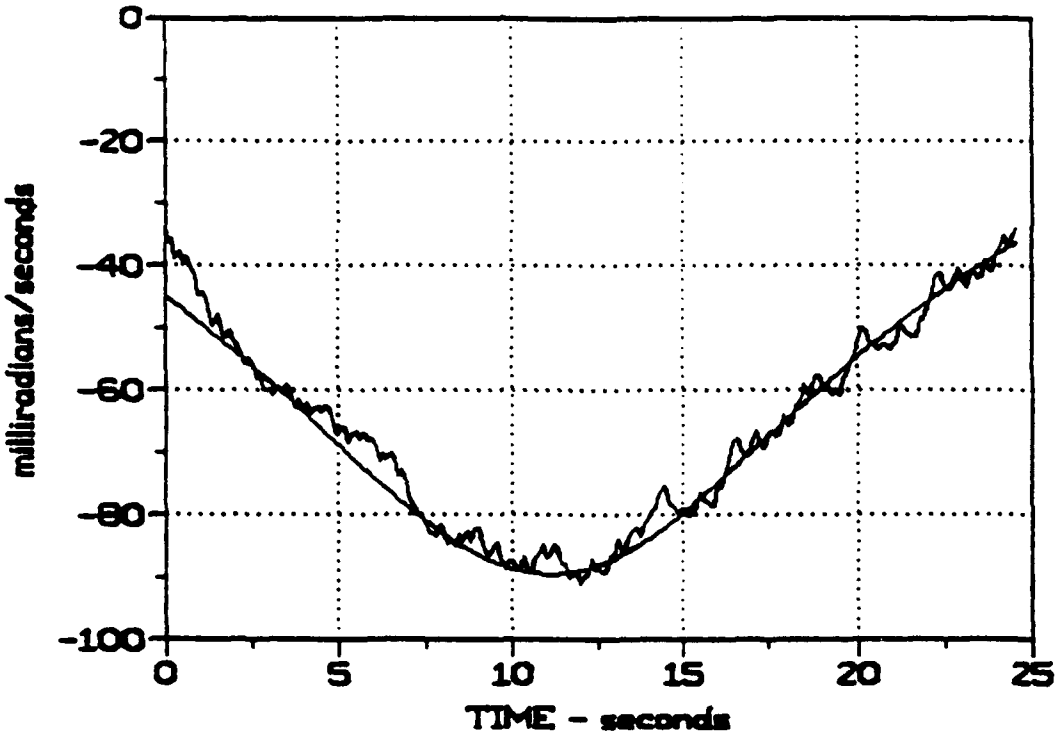


Figure IV-8

Filtered Results - KPA = 0.1, TERM2 = 0.9,
GXHAT(0) = 1st Measurement, Twice Sampling Rate

COMPUTER PROGRAM: FILTER.BAS

PURPOSE: Computes the Filtered Azimuth Rate. Used in Optimization of Filter Coefficients and Investigating Effect of Increasing the Sampling Rate

RESULTS: Figures IV-5 through IV-7

```
010 OPEN "I", #1, "C:RAWDATA.DAT"
20 OPEN "O", #2, "C:AZRATE.CS4"
30 KPA = .1
40 KPE = .1
50 KPR = .1
60 KVR = .1
70 C1 = .9
80 TERM1 = .9
90 TERM2 = .9
100 DTR = .22
110 C2 = DTR*C1
120 C3 = 1! - KVR*DTR
130 INPUT #1, TIME, RSL, SENRNG, AZRATE, GXHAT
140 R(1) = SENRNG
150 RHAT = SENRNG
160 AZDHAT = GXHAT/1000
170 GOSUB 370
180 INPUT #1, TIME, RSL, SENRNG, AZRATE, GXHAT
190 R(2) = SENRNG
200 RHAT = .5*(R(2) + RHAT)
210 RHAT = .5*(SENRG - RHAT)/DTR
220 GOSUB 400
230 GOSUB 370
240 INPUT #1, TIME, RSL, SENRNG, AZRATE, GXHAT
250 RHAT = (R(1) + R(2) + SENRNG)/3!
260 RHATD = .5*(SENRNG - R(1))/DTR
270 GOSUB 400
280 GOSUB 370
290 INPUT #1, TIME, RSL, SENRNG, AZRATE, GXHAT
300 RHAT = RHAT*C1 + RHATD*C2 +SENRNG*KPR
310 TERM3 = (ELDHAT^2 + AZDHAT^2)*DTR
320 RHATD = RHAT*(TERM3 - KVR) + RHATD*C3 +SENRNG*KVR
330 GOSUB 400
340 GOSUB 370
350 GOTO 290
360 END
370 PRINT TIME, RHAT, AZDHAT*1000
380 PRINT #2, TIME, AZDHAT*1000
390 RETURN
400 AZDHAT = TERM2*(1! - 2!*RHATD*DTR/RHAT)*AZDHAT + KPA*GXHAT/1000
410 RETURN
```

COMPUTER PROGRAM: RAWDATPR.BAS

PURPOSE: Computes the Slant Rate and Azimuth Rate for a Given Flight Trajectory. Adds to Generic Noise Theoretical Range and Azimuth Rate. Used for Input for Filter Optimization Program - FILTER.BAS

RESULTS: Figure IV-4

```
010 PI = 3.1416
20 OPEN "I", #1, "C:NOISE.DAT"
30 OPEN "O", #2, "C:AZRATE.THE"
40 OPEN "O", #3, "C:AZNOISE.ACT"
50 OPEN "O", #4, "C:RAWDATA.DAT"
60 INPUT "Aircraft Velocity (knots) = ?", VAC
70 INPUT "Range at Closest Approach = ?", RMIN
80 INPUT "Altitude (feet) = ?", H
90 VAC = VAC*88/60/.8684
100 RMIN = RMIN*3.28
110 RBASE = RMIN - VAC*T
120 RSL = SQR(RMIN^2 + H^2 + RBASE^2)
130 AZ = RBASE/RSL
140 AZ = ATN(AZ/SQR(1 - AZ^2))
150 AZRATE = VAC*RMIN/RSL/RSL
160 INPUT #1, T, AZNOISE, ELNOISE, RNGNOISE
170 AZRATEN = AZRATE*1000 + AZNOISE
180 RNGN = RSL + RNGNOISE
190 PRINT T, RSL, RNGN, -AZRATE*1000, -AZRATEN
200 PRINT #4, T, RSL, RNGN, -AZRATE*1000, -AZRATEN
210 PRINT #2, T, -AZRATE*1000
220 PRINT #3, T, -AZRATEN
230 IF T > 24.5 GOTO 250
240 GOTO 110
250 END
```

CHAPTER V

MEASUREMENTS ALONG GUN BORESIGHT VERSUS LASER BEAM AXIS

In the current BSTING system, the azimuth and elevation rates are measured along the boresight axis though computation of corrections are based on measurements along the laser beam axis. To compensate for this difference, measurements are made when the gun is locked to the laser beam axis. In Chapter II a method for correcting for the difference in measurement axis was proposed. This would allow the gun to be unlocked from the laser beam while the measurements are made, a requirement if a continuous aiming correction is to be achieved.

An investigation was made to determine the aiming error that results if the gun is "unlocked" from the laser beam and the azimuth and elevation rate sensors move with the gun when making measurements. Thus these sensors are not only measuring the angular rates of the aircraft with respect to the target, but also the rate-of-change of the azimuth and elevation corrections. This results from the fact that new azimuth and

elevation corrections are constantly being made. Since the range and range rate measurements are made along the LOS, i.e. the laser beam, they are not affected. The object here was to determine the magnitude of the error in aiming that results because azimuth and elevation rates are made along the gun barrel axis and not the LOS.

The technique used here included the following steps:

a. The "theoretical" azimuth and elevation rates were generated assuming that the laser beam was aimed at the target.

b. The "theoretical" azimuth and elevation corrections needed to keep the beam on target when the gun is used for aiming was calculated.

c. Next the "unlocked" gun azimuth rates were computed using the relationship where AZ_{corr} is constantly changing.

$$\dot{AZ}_{gun} = \frac{V_{a/c} \; Cos(AZ_{laser} + AZ_{corr})}{RSL} \qquad (V\text{-}1)$$

d. For the elevation rate of the gun, first the elevation of the laser beam during target tracking was computed at each measurement interval.

$$EL_{laser} = Arcsin(Altitude/RSL) \qquad (V\text{-}2)$$

The elevation rate was then determined by differencing elevation angles over succeeding measurement intervals, i.e.:

$$ELDOT_{laser} = [EL_{laser}(t + dt) - EL_{laser}(t)]/dt \qquad (V\text{-}3)$$

The elevation correction based on the LOS rate is then

$$EL_{corr} = RSL \times ELDOT_{laser} / V_{muz} \qquad (V\text{-}4)$$

e. The gravity drop correction was then added to this correction. The gravity drop correction was based on the algorithms used in the current BSTING system and is detailed in table V-1.

TABLE V-1

BSTING GRAVITY DROP CORRECTION

$$EL = 16.85 \times (BETA \times TOF)^2 / RSL$$

Where:

$$TOF = RSL/(1 - C*RSL/4)^2 / VP$$

$$VP = Range\ Rate + [V_{muz} \times Cos(EL_{corr}) Cos(AZ_{corr})]$$

$$C = 1.5 \times density/WCDA$$

$$WCDA = 5E\text{-}6\ RSL^2 - 0.036938\ RSL + 700.4299$$

$$BETA = 3E\text{-}9\ RSL^2 - 4.706E\text{-}5\ RSL + 1.0216$$

$$RHO = 0.076479$$

f. The total elevation of the gun becomes

$$EL_{gun} = EL_{laser} + (EL_{corr} + EL_{gd}) \qquad (V\text{-}5)$$

g. The gun elevation rate can be determined again by differencing:

$$ELDOT_{gun} = [EL_{gun}(t + dt) - EL_{gun}(t)]/dt \qquad (V\text{-}6)$$

h. Noise was added to the range and azimuth/elevation measurements to simulate the gunner's ability to keep the laser beam on the target. Again this was based on data obtained during the BSTING flight test program.

i. Since the range data is measured only along the laser beam axis, the same noisy range data was used for both cases.

The five scenarios used in this investigation are summarized in table V-2.

TABLE V-2

SCENARIOS

CASE	ALTITUDE (feet)	VELOCITY (knots)	MINIMUM RANGE (meters)
1	250	150	1000
2	250	175	2000
3	250	100	1500
4	250	75	500
5	250	200	1000

The same noise was used in the analysis of each scenario. Again this was based on noise extracted from data obtained during the BSTING flight test program and represents the gunner's ability to keep the laser spot aimed on the target. Figure V-1 shows typical noisy azimuth and elevation rate measurements as were used in the subsequent analysis. These examples are for measurement along laser beam axis. Noise was added to the rates measured along gun boresight axis.

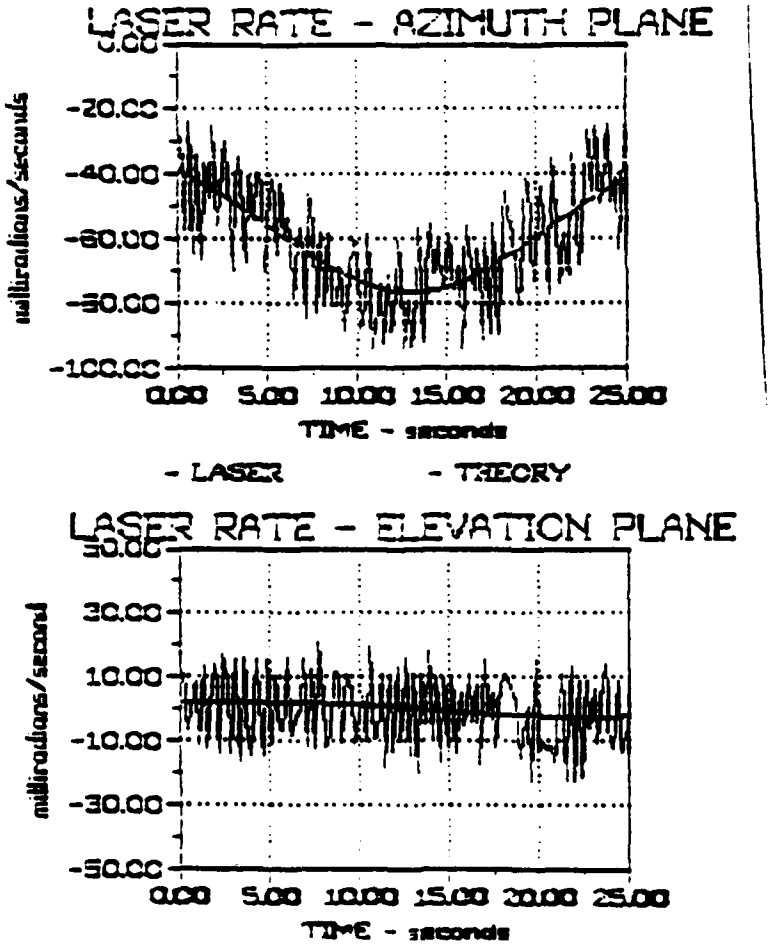


Figure V-1

Representive Azimuth and Elevation Rate Measurement
(Case 1 Shown - Laser Beam Axis)

The investigation shows a significant difference in the time it takes to reach the point where the algorithms provide an aiming solution within the +/- 4 milliradian criteria. See Table V-3. This parameter is important for a "snap and shoot" fire control system. In each case the results are based on azimuth/elevation rates measured along the gun boresight axis.

TABLE V-3

SETTLING TIMES

SCENARIO	TIME TO REACH +/- 4 mr CRITERIA AZIMUTH (seconds)	ELEVATION
1	2	1.5
2	13	21
3	15	2
4	2	1
5	2	2

The results of using rate sensors on the laser axis versus on the gunsight axis are shown in figures V-2 to V-6 for the azimuth correction and figures V-7 to V-11 for the elevation correction. In most cases the gunsight based corrections are not that different from laser based corrections. In some cases, they were actually better.

Table V-4 shows the maximum aiming errors (After the +/- 4 criteria was reached) for laser and boresight based sensor measurements.

TABLE V-4

MAXIMUM ERROR IN AIMING CORRECTIONS

CASE	MAXIMUM ERROR (milliradians) MEASUREMENT AXIS			
	LASER BEAM		GUN BORESIGHT	
	AZIMUTH	ELEVATION	AZIMUTH	ELEVATION
1	3	4	3	4
2	4	3	5	7
3	3	2	8	3.5
4	1.5	3	1	3
5	4.5	4	3	3

Again it appears that rate measurements made along the boresight do not result in aiming errors much different those that would be made along the laser beam axis.

LASER BEAM AXIS

GUN BORESIGHT

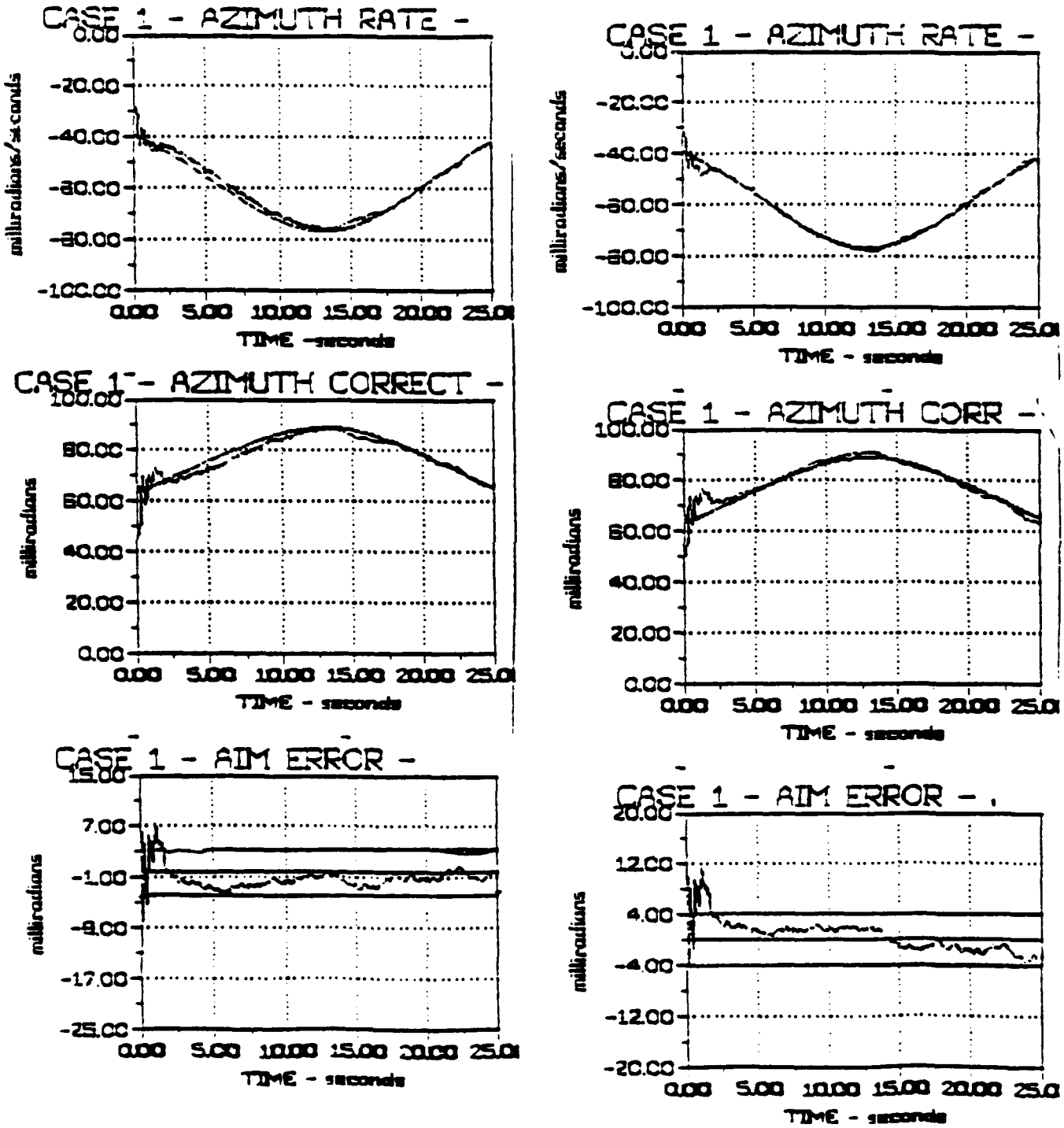


Figure V-2

Altitude = 250 ft, Velocity = 150 kts, Minimum Range = 1000 mtr

LASER BEAM AXIS

GUN BORESIGHT

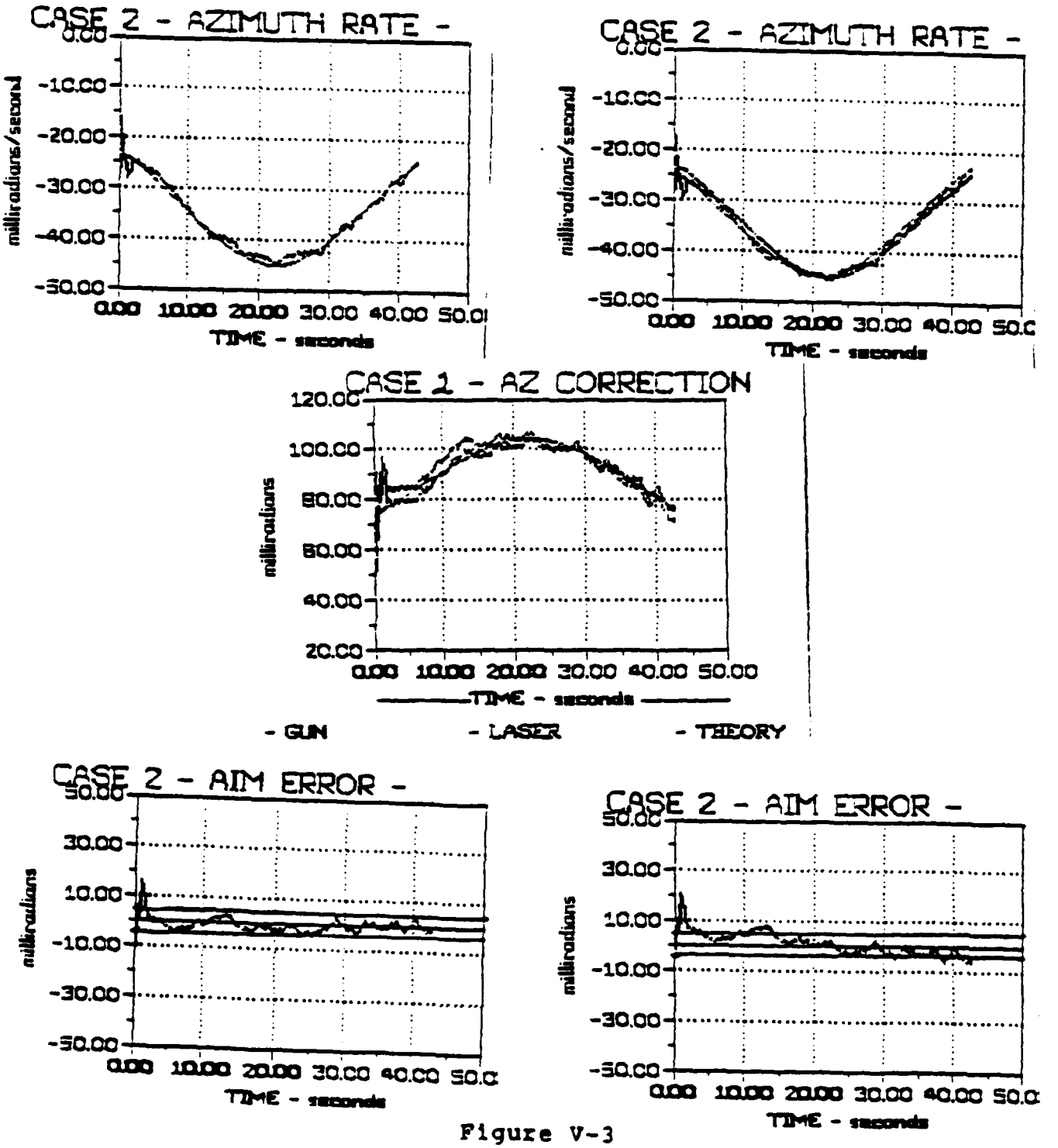


Figure V-3

Altitude = 250 ft, Velocity = 175 kts, Minimum Range = 2000 mtr

LASER BEAM AXIS

GUN BORESIGHT

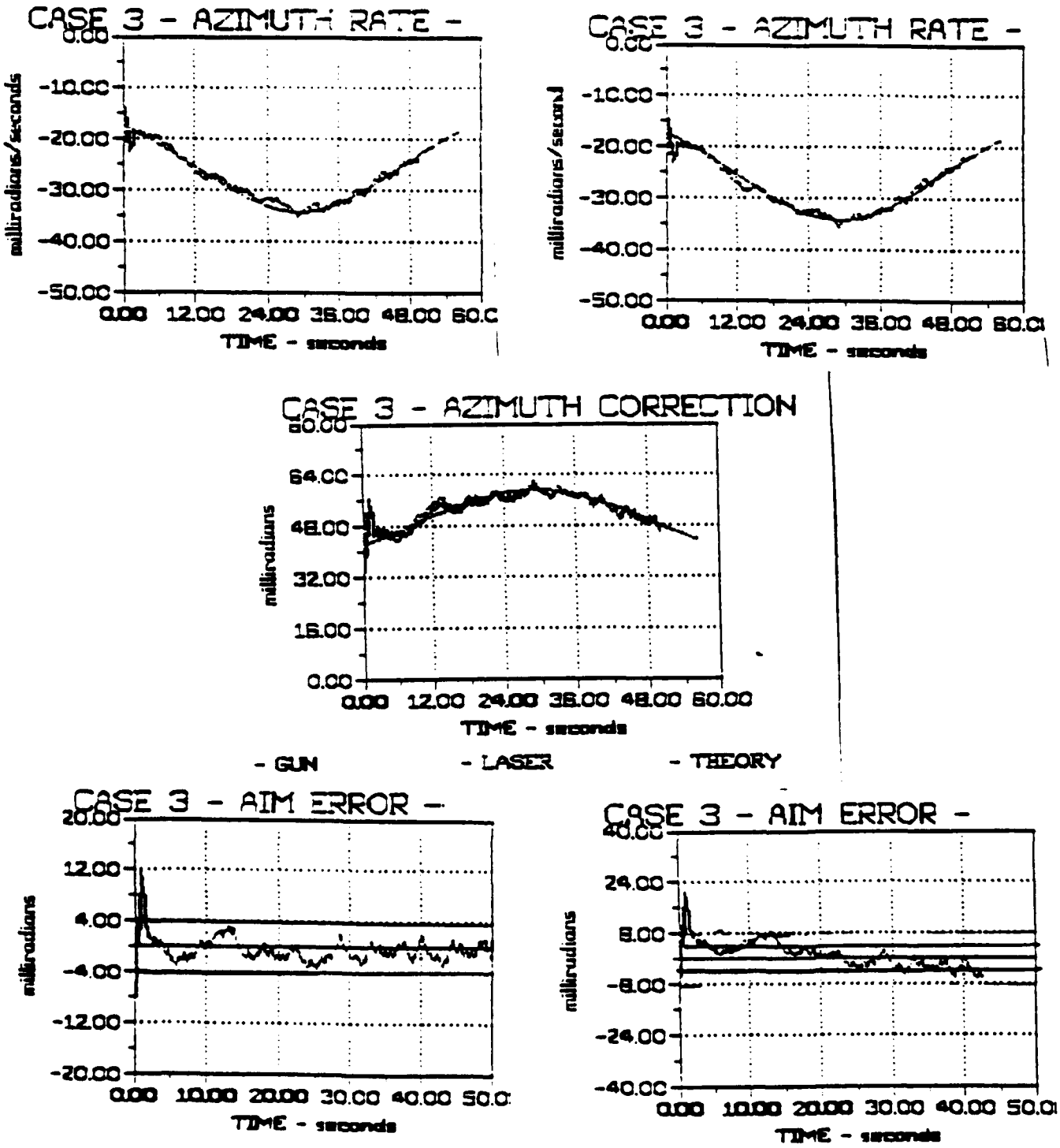


Figure V-4

Altitude = 250 ft, Velocity = 100 kts, Minimum Range = 1500 mtr

LASER BEAM AXIS

GUN BORESIGHT

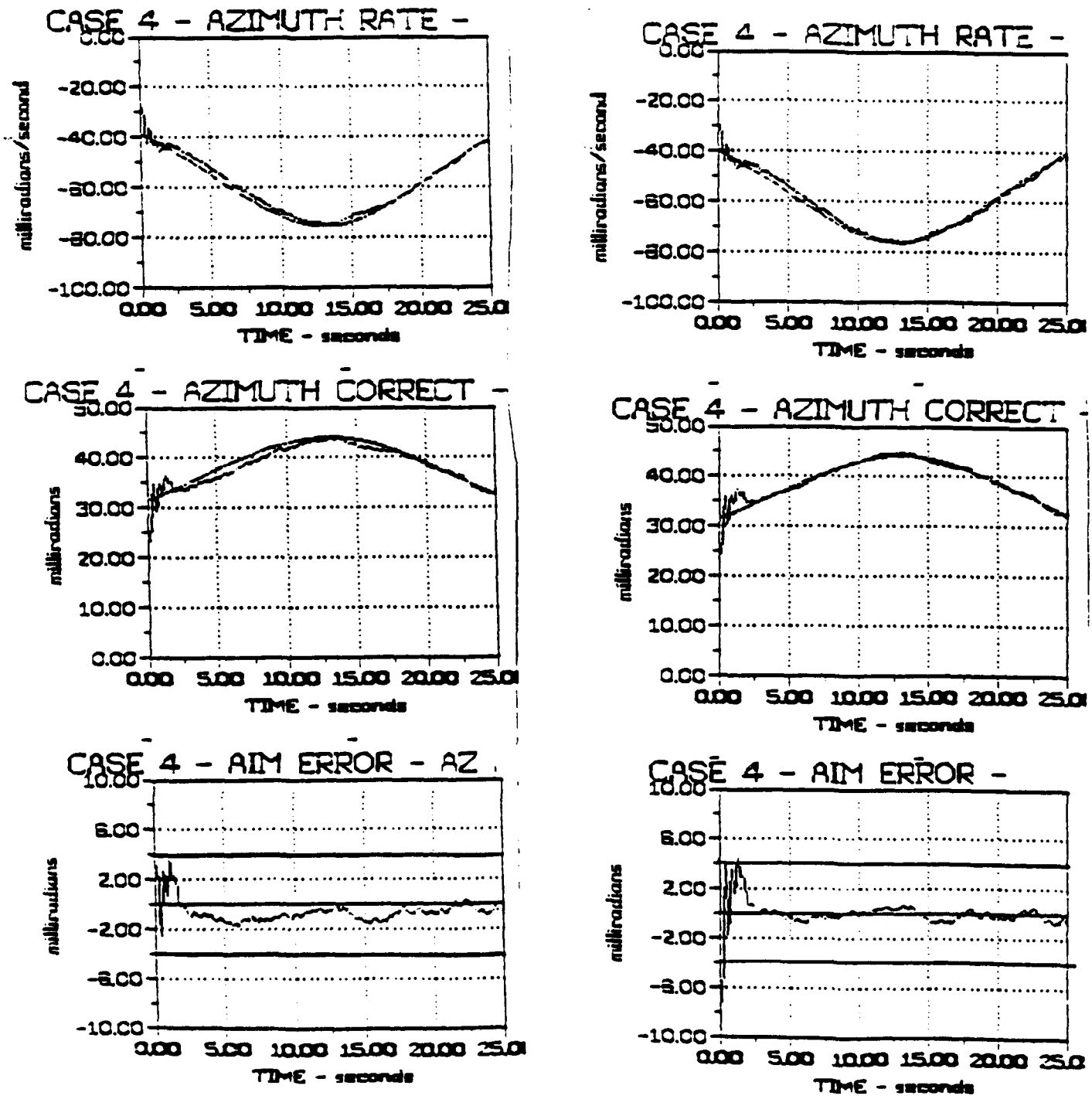


Figure V-5

Altitude = 250 ft, Velocity = 75 kts, Minimum Range = 500 mtr

LASER BEAM AXIS

GUN BORESIGHT

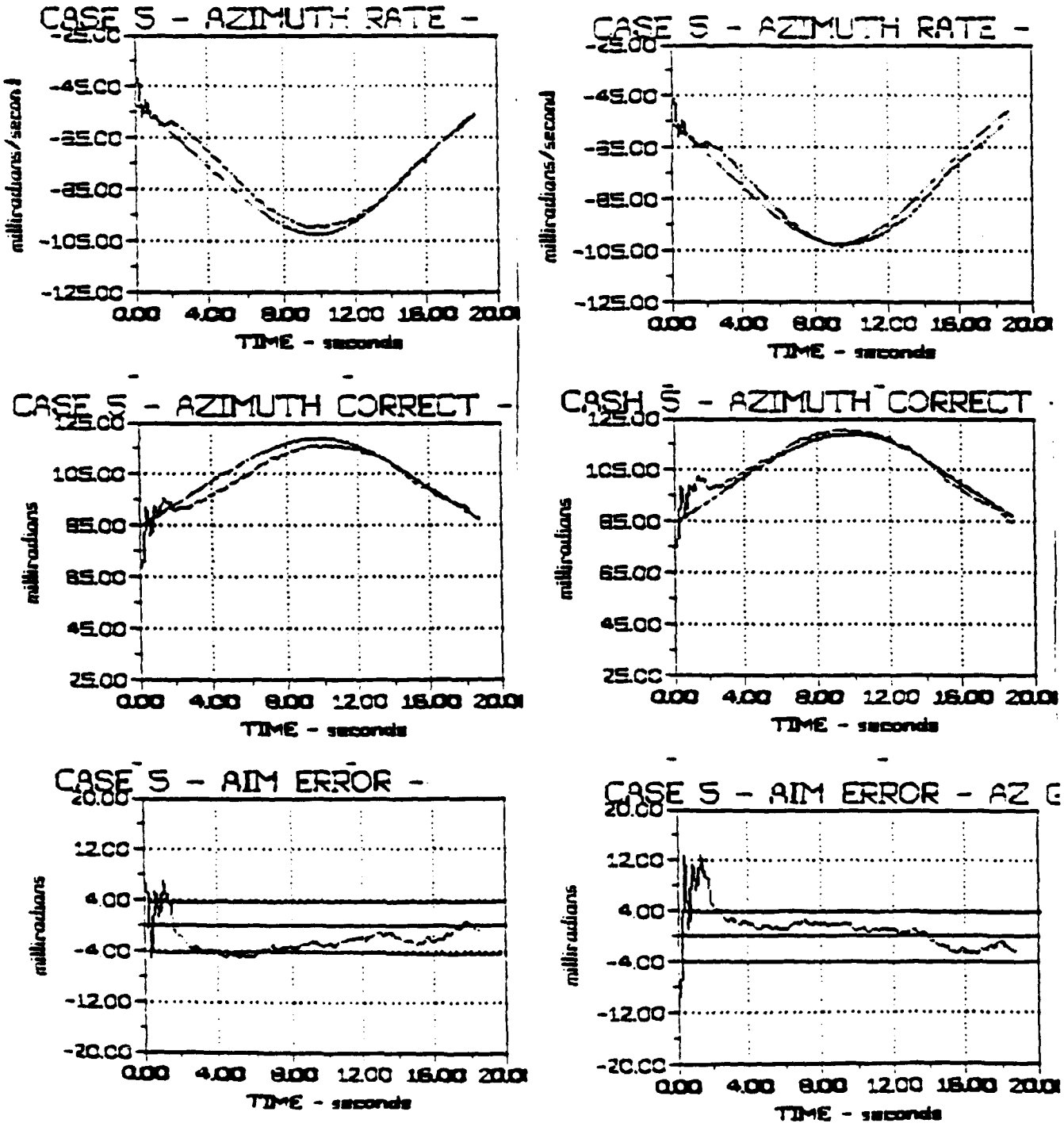


Figure V-6

Altitude = 250 ft, Velocity = 200 kts, Minimum Range = 1000 mtr

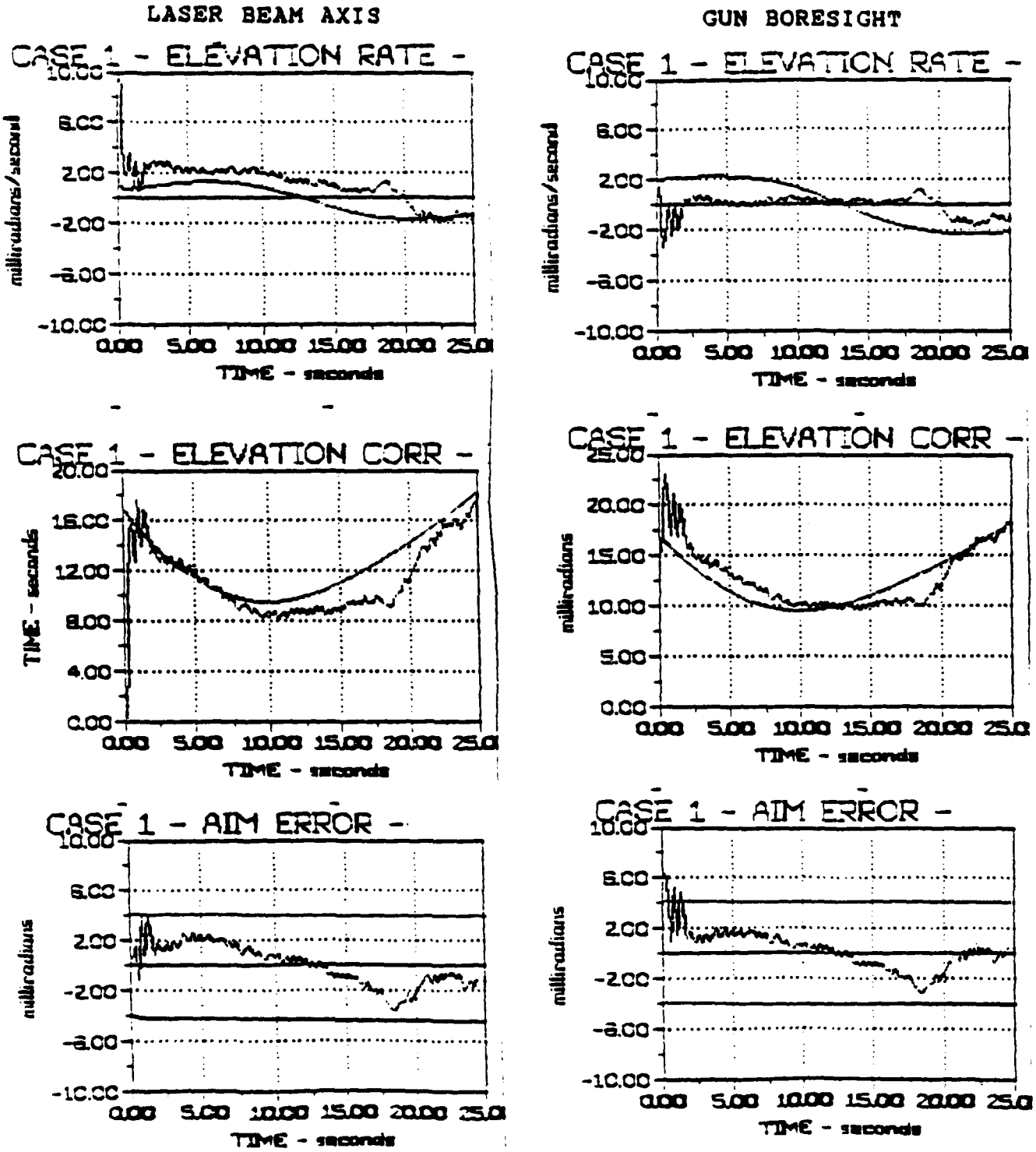


Figure V-7

Altitude = 250 ft, Velocity = 150 kts, Minimum Range = 1000 mtr

LASER BEAM AXIS

GUN BORESIGHT

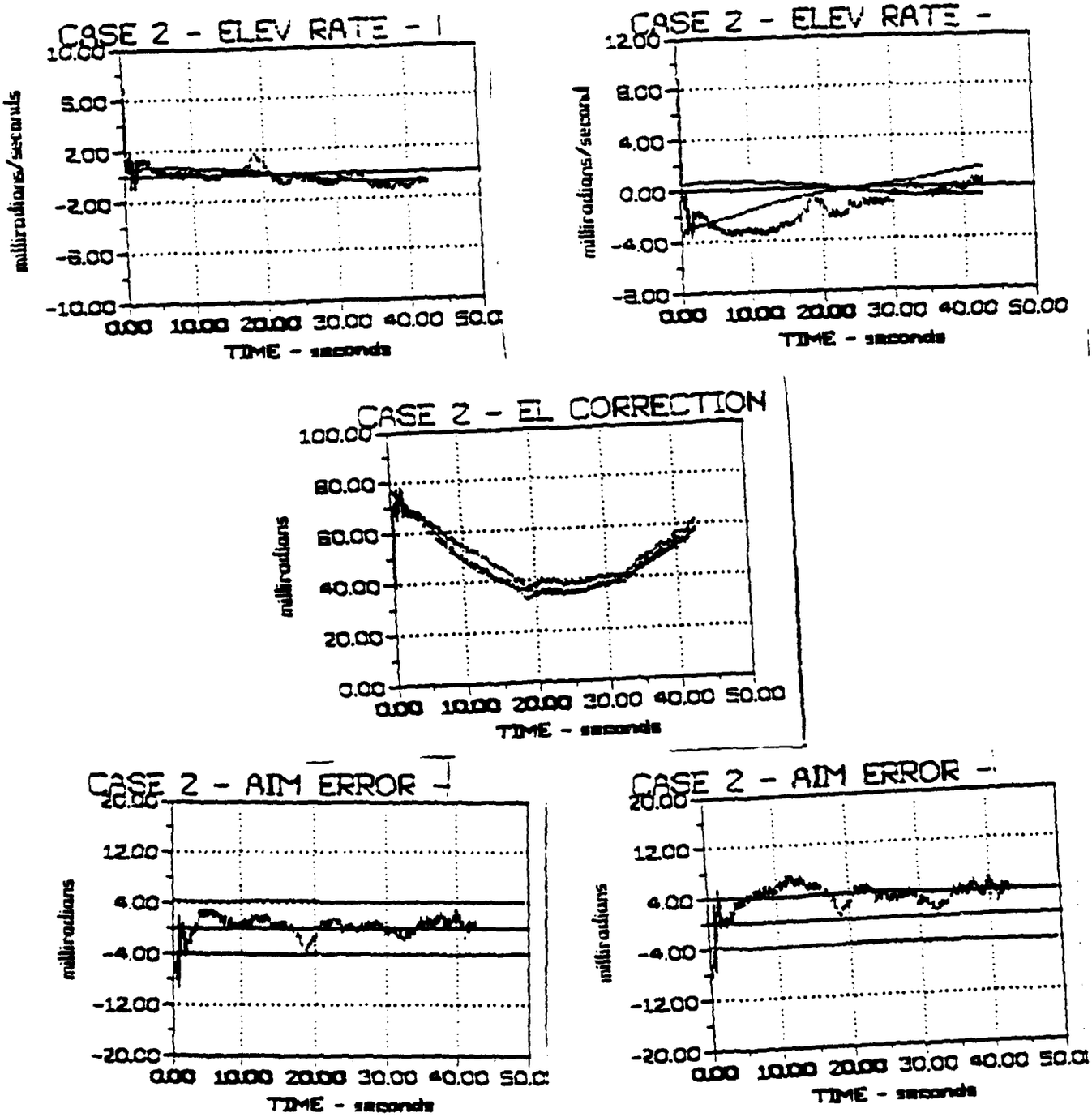


Figure V-8

Altitude = 250 ft, Velocity = 175 kts, Minimum Range = 2000 mtr

LASER BEAM AXIS

GUN BORESIGHT

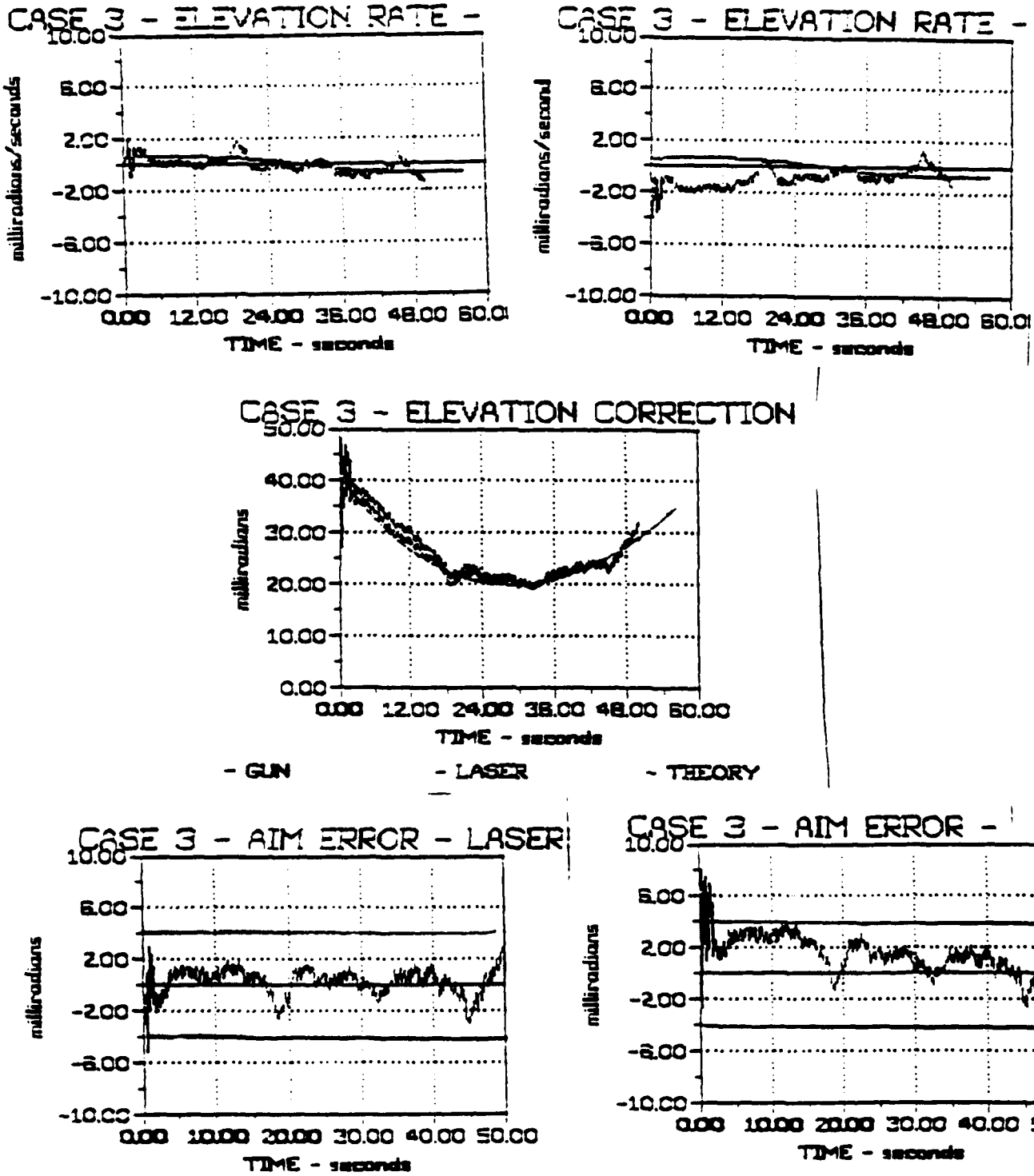


Figure V-9

Altitude = 250 ft, Velocity = 100 kts, Minimum Range = 1500 mtr

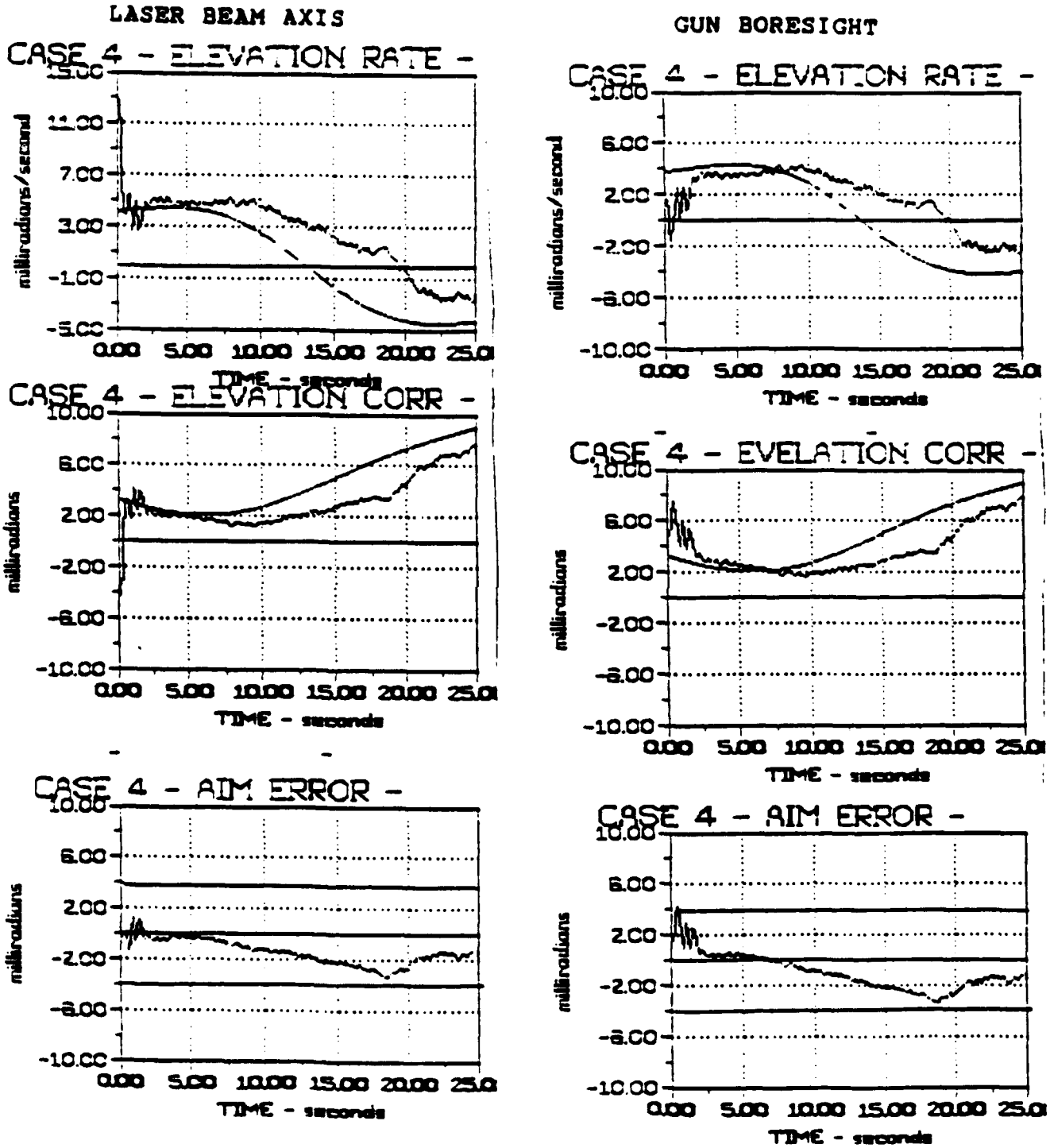


Figure V-10

Altitude = 250 ft, Velocity = 75 kts, Minimum Range = 500 mtr

LASER BEAM AXIS

GUN BORESIGHT

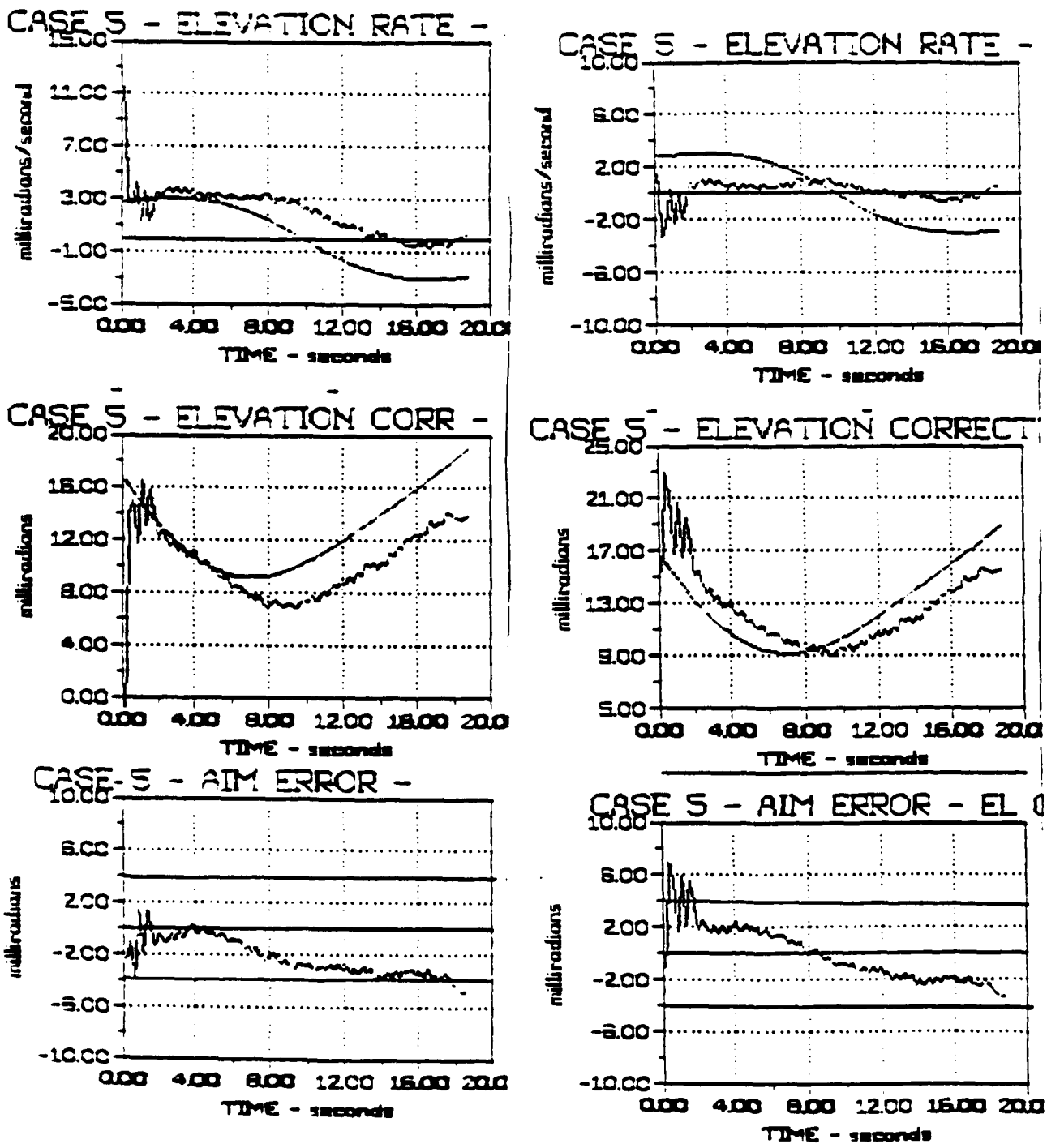


Figure V-11

Altitude = 250 ft, Velocity = 200 kts, Minimum Range = 1000 mtr

COMPUTER PROGRAM: LASERPR.BAS

PURPOSE: Computes the Theoretical Range and Azimuth/Elevation Rates along the Laser Beam Axis.

RESULTS: Used as Input LASNOPR.BAS

```
010 OPEN "O", #1, "B:LASAZCO1.out"
20 OPEN "O", #2, "B:LASAZDT1.OUT"
30 OPEN "O", #3, "B:LASELDT1.OUT"
40 OPEN "O", #4, "B:LASDATA1.OUT"
50 OPEN "O", #5, "B:LASELCO1.OUT"
60 INPUT "Altitude (feet) = ?", H
70 INPUT "Aircraft Velocity (knots) =?", VAC
80 VAC = VAC*88/60/.8684
90 INPUT "Range at Closest Approach (meters) = ?", RMIN
100 RMIN = RMIN*3.28
110 RBASE = RMIN - VAC*T
120 RSL = SQR(H^2 + RMIN^2 + RBASE^2)
130 RDOT = -RBASE*VAC/RSL
140 AZDOT = -RMIN*VAC/RSL/RSL
150 AZ = ATN(RBASE/RMIN)*180/3.1416
160 BOTTOM = SQR(RMIN^2 + RBASE^2)
170 EL = ATN(H/BOTTOM)*180/3.1417
180 ELDOT = (EL - EL1)*3.1416/180/.22
190 EL1 = EL
200 ELCORR = RSL*ELDOT/2840
210 ELCORR = -ATN(ELCORR/SQR(1 - ELCORR^2))
220 ELCORR = ELCORR*1000
230 AZCORR = RSL*AZDOT/2840/COS(ELCORR/1000)
240 AZCORR = -ATN(AZCORR/SQR(1 - AZCORR^2))
250 AZCORR = AZCORR*1000
260 VP = RDOT + 2840*COS(ELCORR/1000)*COS(AZCORR/1000)
270 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
280 BETA =  3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
290 RHO = .076479
300 C = 1.5*RHO/WCDA
310 TOF = RSL/VP/(1 - C*RSL/4)^2
320 ELGD = 16.085*(BETA*TOF)^2/RSL
330 ELGD = ELGD*1000
335 ELCORR = ELCORR + ELGD
340 PRINT #1, T, AZCORR
350 PRINT #2, T, AZDOT*1000
360 PRINT #3, T, ELDOT*1000
370 PRINT T, RSL, AZ, EL, AZDOT*1000, ELDOT*1000, AZCORR, ELCORR, RDOT, ELGD
380 PRINT #4, T, RSL, AZ, EL, AZDOT*1000, ELDOT*1000, AZCORR, ELCORR, RDOT, ELGD
390 PRINT #5, T, ELCORR
```

```
400 IF AZ < -42.6 THEN GOTO 430
410 T = T + .22
420 GOTO 110
430 END
```

COMPUTER PROGRAM: GUNPR.BAS

PURPOSE: Computes the Theoretical Range and Azimuth/Elevation Rates along the Gun Boresight

RESULTS: Used as Input to GUNNOPR.BAS

```
  10 OPEN "O", #1, "B:GUNAZ1.out"
 20 OPEN "O", #2, "B:GUNAZDT1.OUT"
 30 OPEN "O", #3, "B:GUNELDT1.OUT"
 40 OPEN "O", #4, "B:GUNDATA1.OUT"
 50 OPEN "O", #5, "B:GUNEL1.OUT"
 60 INPUT "Altitude (feet) = ?", H
 70 INPUT "Aircraft Velocity (knots) =?", VAC
 80 VAC = VAC*88/60/.8684
 90 INPUT "Range at Closest Approach (meters) = ?", RMIN
 100 RMIN = RMIN*3.28
 110 RBASE = RMIN - VAC*T
 120 RSL = SQR(H^2 + RMIN^2 + RBASE^2)
 130 RDOT = -RBASE*VAC/RSL
 140 BOTTOM = SQR(RMIN^2 + RBASE^2)
 150 EL = ATN(H/BOTTOM)
 160 ELDOT = (EL - EL1)/.22
 170 EL1 = EL
 180 ELCORR = RSL*ELDDOT/2840
 190 ELCORR = -ATN(ELCORR/SQR(1 - ELCORR^2))
 200 VP = RDOT + 2840*COS(ELCORR)*COS(AZCORR)
 210 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
 220 BETA =  3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
 230 RHO = .076479
 240 C = 1.5*RHO/WCDA
 250 TOF = RSL/VP/(1 - C*RSL/4)^2
 260 ELGD = 16.085*(BETA*TOF)^2/RSL
 270 ELCORRT = ELGD + ELCORR
 280 ELGUN = EL + ELCORRT
 290 ELDGUN = (ELGUN - ELGUN1)/.22
 300 ELGUN1 = ELGUN
 310 AZ = ATN(RBASE/RMIN)
```

```
320 AZDOT = -RMIN*VAC/RSL/RSL
330 AZCORR = RSL*AZDOT/2840/COS(ELCORR)
340 AZCORR = -ATN(AZCORR/SQR(1 - AZCORR^2))
350 AZDGUN = -VAC*COS(AZ*PI/180 - AZCORR)/RSL
360 AZGUN = AZ + AZCORR
370 PRINT #1, T, AZGUN*180/3.1416
380 PRINT #2, T, AZDGUN*1000
390 PRINT #3, T, ELDGUN*1000
400 PRINT T, EL, ELCORRT, ELGUN
410 PRINT #4, T, AZGUN*180/3.1417, ELGUN*180/3.1416, AZDGUN*1000, ELDGUN*1000
420 PRINT #5, T, ELGUN*180/3.1416
430 AZI = AZ*180/3.1416
440 IF AZI < -42.6 THEN GOTO 470
450 T = T + .22
460 GOTO 110
470 END
```

COMPUTER PROGRAM: LASNOPR.BAS

PURPOSE: Adds Noise to the Theoretical Range and Azimuth/ Elevation Measurements along Laser Beam Axis in LASERPR.BAS

RESULTS: Used as Input to FILTLAS.BAS

```
10 OPEN "I", #1, "B:NOISE.DAT"
20 OPEN "I", #2, "B:LASDATA1.OUT"
30 OPEN "O", #3, "B:NOISYAZ1.LAS"
40 OPEN "O", #4, "B:NOISYEL1.LAS"
50 OPEN "O", #5, "B:NOISYRG1.LAS"
60 INPUT #1, T, AZNOISE, ELNOISE, RNGNOISE
70 INPUT #2, T, RSL, AZ, EL, AZDOT, ELDOT, AZCORR, ELCORR, RDOT, ELGD
80 NOISERSL = RSL + RNGNOISE
90 NOISEAZ = AZDOT + AZNOISE
100 NOISEEL = ELDOT + ELNOISE
110 PRINT #3, T, NOISEAZ
120 PRINT #4, T, NOISEEL
130 PRINT #5, T, NOISERSL
140 PRINT T, NOISEAZ, NOISEEL, NOISERSL
150 GOTO 60
```

COMPUTER PROGRAM: GUNNOPR.BAS

0 PURPOSE: Adds Noise to the Theoretical Range and Azimuth/ Elevation Measurements along Boresight Axis Determined in GUNPR.BAS

RESULTS: Used as Input to FILTGUN.BAS

```
 10 OPEN "I", #1, "B:NOISE.DAT"
20 OPEN "I", #2, "B:GUNDATA1.OUT"
30 OPEN "O", #3, "B:NOISYAZ1.GUN"
40 OPEN "O", #4, "B:NOISYEL1.GUN"
60 INPUT #1, T, AZNOISE, ELNOISE, RNGNOISE
70 INPUT #2, T, AZGUN, ELGUN, AZDOT, ELDOT
90 NOISEAZ = AZDOT + AZNOISE
100 NOISEEL = ELDOT + ELNOISE
110 PRINT #3, T, NOISEAZ
120 PRINT #4, T, NOISEEL
140 PRINT T, NOISEAZ, NOISEEL, NOISERSL
150 GOTO 60
```

```
190 RHAT = SENRNG
200 AZDHAT = GXHAT/1000
210 ELDHAT = GYHAT/1000
220 GOSUB 870
230 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
240 R(2) = SENRNG
250 RHAT = .5*(R(2) + RHAT)
260 RHATD = .5*(SENRNG - RHAT)/DTR
270 AZDHAT = (AZDHAT + GXHAT/1000)/2
280 ELDHAT = (ELDHAT + GYHAT/1000)/2
290 GOSUB 870
300 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
310 R(3) = SENRNG
320 RHAT = (2*RHAT + SENRNG)/3
330 RHATD = .5*(SENRNG - R(2))/DTR
340 AZDHAT = (2*AZDHAT + GXHAT/1000)/3
350 ELDHAT = (2*ELDHAT + GYHAT/1000)/3
360 GOSUB 870
370 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
380 R(4) = SENRNG
390 RHAT = (3*RHAT + SENRNG)/4
400 RHATD = .5*(SENRNG - R(3))/DTR
410 AZDHAT = (3*AZDHAT + GXHAT/1000)/4
420 ELDHAT = (3*ELDHAT + GYHAT/1000)/4
430 GOSUB 870
440 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
450 R(5) = SENRNG
460 RHAT = (4*RHAT + SENRNG)/5
470 RHATD = .5*(SENRNG - R(4))/DTR
480 AZDHAT = (4*AZDHAT + GXHAT/1000)/5
490 ELDHAT = (4*ELDHAT + GYHAT/1000)/5
500 GOSUB 870
510 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
520 R(6) = SENRNG
530 RHAT = (5*RHAT + SENRNG)/6
540 RHATD = .5*(SENRNG - R(5))/DTR
550 AZDHAT = (5*AZDHAT + GXHAT/1000)/6
560 ELDHAT = (5*ELDHAT + GYHAT/1000)/6
570 GOSUB 870
580 INPUT #1, TIME, GXHAT, GYHAT  SENRNG
590 R(7) = SENRNG
600 RHAT = (6*RHAT + SENRNG)/7
610 RHATD = .5*(SENRNG - R(6))/DTR
620 AZDHAT = (6*AZDHAT + GXHAT/1000)/7
630 ELDHAT = (6*ELDHAT + GYHAT/1000)/7
640 GOSUB 870
650 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
660 R(8) = SENRNG
670 RHAT = (7*RHAT + SENRNG)/8
680 RHATD = .5*(SENRNG - R(7))/DTR
690 AZDHAT = (7*AZDHAT + GXHAT/1000)/8
700 ELDHAT = (7*ELDHAT + GYHAT/1000)/8
```

```
710 GOSUB 870
720 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
730 R(9) = SENRNG
740 RHAT = (8*RHAT + SENRNG)/9
750 RHATD = .5*(SENRNG - R(8))/DTR
760 AZDHAT = (8*AZDHAT + GXHAT/1000)/9
770 ELDHAT = (8*ELDHAT + GYHAT/1000)/9
780 GOSUB 870
790 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
800 RHAT = RHAT*C1 + RHATD*C2 +SENRNG*KPR
810 TERM3 = (ELDHAT^2 + AZDHAT^2)*DTR
820 RHATD = RHAT*(TERM3 - KVR) + RHATD*C3 +SENRNG*KVR
830 GOSUB 1050
840 GOSUB 870
850 GOTO 790
860 END
870 ELCORR = RHAT*ELDHAT/2840
880 ELCORR = -ATN(ELCORR/SQR(1 - ELCORR^2))
890 AZCORR = RHAT*AZDHAT/2840
900 AZCORR = -ATN(AZCORR/(1 - AZCORR^2))
920 VP = RHATD + 2840*COS(ELCORR)*COS(AZCORR)
930 WCDA = .000005*RHAT^2 - .036938*RHAT + 700.429
940 BETA = 3E-09*RHAT^2 - 4.706E-05*RHAT + 1.0216
950 RHO = .076479
960 C = 1.5*RHO/WCDA
970 TOF = RHAT/VP/(1 - C*RHAT/4)^2
980 ELGB = 16.085*(BETA*TOF)^2/RHAT
981 ELCORR = ELCORR + ELGB
990 PRINT TIME, AZDHAT*1000, ELDHAT*1000, AZCORR*1000, ELCORR*1000
1000 PRINT #2, TIME, AZDHAT*1000
1010 PRINT #3, TIME, ELDHAT*1000
1020 PRINT #4, TIME, AZCORR*1000
1030 PRINT #5, TIME, ELCORR*1000
1040 RETURN
1050 ELDHAT = TERM1*(1! - 2!*RHATD*DTR/RHAT)*ELDHAT + KPE*GYHAT/1000
1060 AZDHAT = TERM2*(1! - 2!*RHATD*DTR/RHAT)*AZDHAT + KPA*GXHAT/1000
1070 RETURN
1080 END
```

COMPUTER PROGRAM: FILTLAS.BAS

PURPOSE: Computes the Azimuth and Elevation Aiming Corrections from Noisy Sensor Data Measured along Laser Beam Axis. Used Data from LASNOPR.BA

RESULTS: Figures of Azimuth/Elevation Rates and Azimuth/Elevation Aiming Corrections - Laser Axis

```
10 OPEN "I", #1, "B:NOISE1.LAS"
20 OPEN "O", #2, "B:AZDOT1.LAS"
30 OPEN "O", #3, "B:ELDOT1.LAS"
40 OPEN "O", #4, "B:AZCORR1N.LAS"
50 OPEN "O", #5, "B:ELCORR1N.LAS"
70 KPA = .1
80 KPE = .1
90 KPR = .1
100 KVR = .1
110 C1 = .9
120 TERM1 = .9
130 TERM2 = .9
140 DTR = .22
150 C2 = DTR*C1
160 C3 = 1! - KVR*DTR
170 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
180 R(1) = SENRNG
190 RHAT = SENRNG
200 AZDHAT = GXHAT/1000
210 ELDHAT = GYHAT/1000
220 GOSUB 870
230 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
240 R(2) = SENRNG
250 RHAT = .5*(R(2) + RHAT)
260 RHATD = .5*(SENRNG - RHAT)/DTR
270 AZDHAT = (AZDHAT + GXHAT/1000)/2
280 ELDHAT = (ELDHAT + GYHAT/1000)/2
290 GOSUB 870
300 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
310 R(3) = SENRNG
320 RHAT = (2*RHAT + SENRNG)/3
330 RHATD = .5*(SENRNG - R(2))/DTR
340 AZDHAT = (2*AZDHAT + GXHAT/1000)/3
350 ELDHAT = (2*ELDHAT + GYHAT/1000)/3
360 GOSUB 870
370 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
```

```
380 R(4) = SENRNG
390 RHAT = (3*RHAT + SENRNG)/4
400 RHATD = .5*(SENRNG - R(3))/DTR
410 AZDHAT = (3*AZDHAT + GXHAT/1000)/4
420 ELDHAT = (3*ELDHAT + GYHAT/1000)/4
430 GOSUB 870
440 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
450 R(5) = SENRNG
460 RHAT = (4*RHAT + SENRNG)/5
470 RHATD = .5*(SENRNG - R(4))/DTR
480 AZDHAT = (4*AZDHAT + GXHAT/1000)/5
490 ELDHAT = (4*ELDHAT + GYHAT/1000)/5
500 GOSUB 870
510 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
520 R(6) = SENRNG
530 RHAT = (5*RHAT + SENRNG)/6
540 RHATD = .5*(SENRNG - R(5))/DTR
550 AZDHAT = (5*AZDHAT + GXHAT/1000)/6
560 ELDHAT = (5*ELDHAT + GYHAT/1000)/6
570 GOSUB 870
580 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
590 R(7) = SENRNG
600 RHAT = (6*RHAT + SENRNG)/7
610 RHATD = .5*(SENRNG - R(6))/DTR
620 AZDHAT = (6*AZDHAT + GXHAT/1000)/7
630 ELDHAT = (6*ELDHAT + GYHAT/1000)/7
640 GOSUB 870
650 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
660 R(8) = SENRNG
670 RHAT = (7*RHAT + SENRNG)/8
680 RHATD = .5*(SENRNG - R(7))/DTR
690 AZDHAT = (7*AZDHAT + GXHAT/1000)/8
700 ELDHAT = (7*ELDHAT + GYHAT/1000)/8
710 GOSUB 870
720 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
730 R(9) = SENRNG
740 RHAT = (8*RHAT + SENRNG)/9
750 RHATD = .5*(SENRNG - R(8))/DTR
760 AZDHAT = (8*AZDHAT + GXHAT/1000)/9
770 ELDHAT = (8*ELDHAT + GYHAT/1000)/9
780 GOSUB 870
790 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
800 RHAT = RHAT*C1 + RHATD*C2 +SENRNG*KPR
810 TERM3 = (ELDHAT^2 + AZDHAT^2)*DTR
820 RHATD = RHAT*(TERM3 - KVR) + RHATD*C3 +SENRNG*KVR
830 GOSUB 1050
840 GOSUB 870
850 GOTO 790
860 END
870 ELCORR = RHAT*ELDHAT/2840
880 ELCORR = -ATN(ELCORR/SQR(1 - ELCORR^2))
890 AZCORR = RHAT*AZDHAT/2840
900 AZCORR = -ATN(AZCORR/(1 - AZCORR^2))
```

```
920 VP = RHATD + 2840*COS(ELCORR)*COS(AZCORR)
930 WCDA = .000005*RHAT^2 - .036938*RHAT + 700.429
940 BETA = 3E-09*RHAT^2 - 4.706E-05*RHAT + 1.0216
950 RHO = .076479
960 C = 1.5*RHO/WCDA
970 TOF = RHAT/VP/(1 - C*RHAT/4)^2
980 ELGB = 16.085*(BETA*TOF)^2/RHAT
981 ELCORR = ELCORR + ELGB
990 PRINT TIME, AZDHAT*1000, ELDHAT*1000, AZCORR*1000, ELCORR*1000
1000 PRINT #2, TIME, AZDHAT*1000
1010 PRINT #3, TIME, ELDHAT*1000
1020 PRINT #4, TIME, AZCORR*1000
1030 PRINT #5, TIME, ELCORR*1000
1040 RETURN
1050 ELDHAT = TERM1*(1! - 2!*RHATD*DTR/RHAT)*ELDHAT + KPE*GYHAT/1000
1060 AZDHAT = TERM2*(1! - 2!*RHATD*DTR/RHAT)*AZDHAT + KPA*GXHAT/1000
1070 RETURN
1080 END
```

COMPUTER PROGRAM: FILTGUN.BAS

PURPOSE: Computes the Azimuth and Elevation Aiming Corrections from Noisy Sensor Data Measured along Gun Boresight Used Data from GUNNOPR.BAS

RESULTS: Figures of Azimuth/Elevation Rates and Azimuth/Elevation Aiming Corrections - Boresight Axis

```
010 OPEN "I", #1, "B:NOISE1.GUN"
20 OPEN "O", #2, "B:AZDOT1.GUN"
30 OPEN "O", #3, "B:ELDOT1.GuN"
40 OPEN "O", #4, "B:AZCORR1N.GUN"
50 OPEN "O", #5, "B:ELCORR1N.GUN"
70 KPA = .03
80 KPE = .03
90 KPR = .1
100 KVR = .1
110 C1 = .9
120 TERM1 = .97
130 TERM2 = .97
140 DTR = .22
150 C2 = DTR*C1
160 C3 = 1! - KVR*DTR
170 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
180 R(1) = SENRNG
```

CHAPTER VI

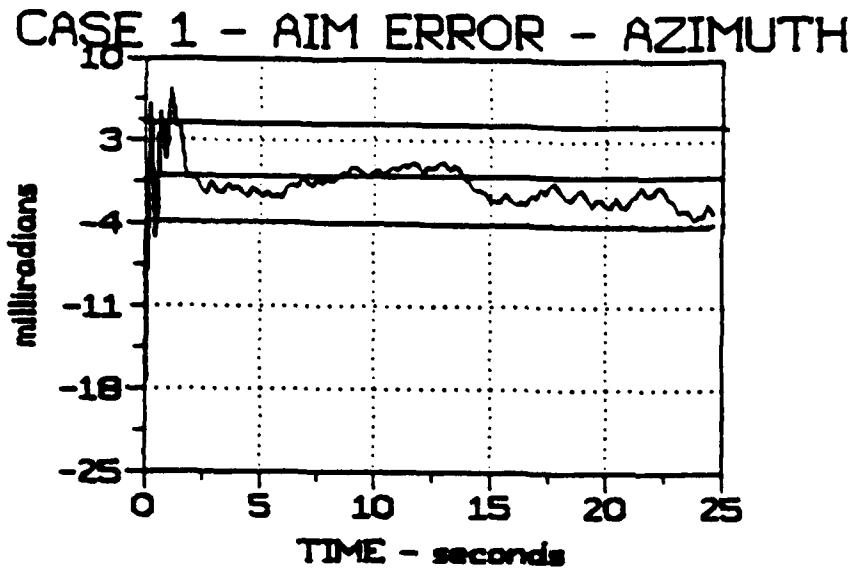
GUN-BASED VERSUS LASER-BASED MEASUREMENTS

Investigation of Initial Conditions

In the previous chapters, it was found that aiming accuracy was basically no worse, or no better either, with measurements made along gun boresight than when they were made along the laser axis. The gun was unlocked from the laser beam for the entire time the gunner was aiming at the target during this analysis.

Next investigated was the effect of "locking" the gun to the laser beam during the first two seconds while the initial solution is computed. This follows the technique used in the current BSTING system. After the initial computation was determined then the gun was "unlocked" so measurements could be used to provide a continuously updated solution using these computed aiming corrections as initial conditions.

The results are shown in figure V-1 through V-5. In these cases, the same five scenarios were used as reported in the previous chapter. There was no significant change in the aiming errors and in most cases the system could meet the +/- 4 milliradians criteria.



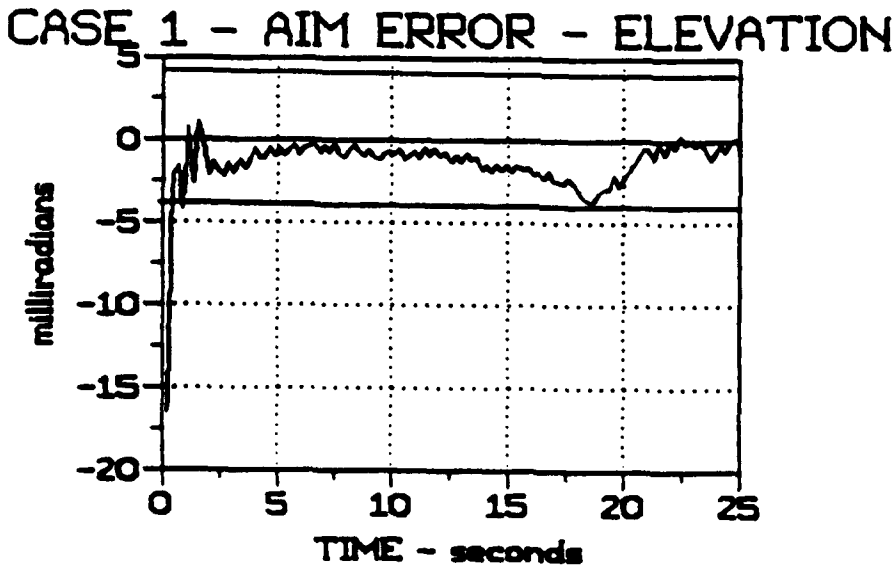
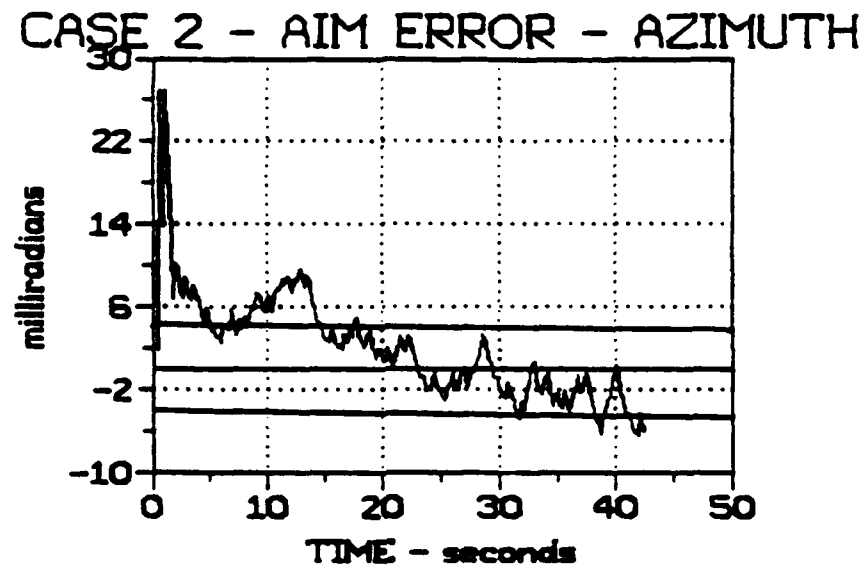


Figure VI-1

Altitude = 250 ft, Velocity = 150 kts, Minimum Range = 1000 mtr



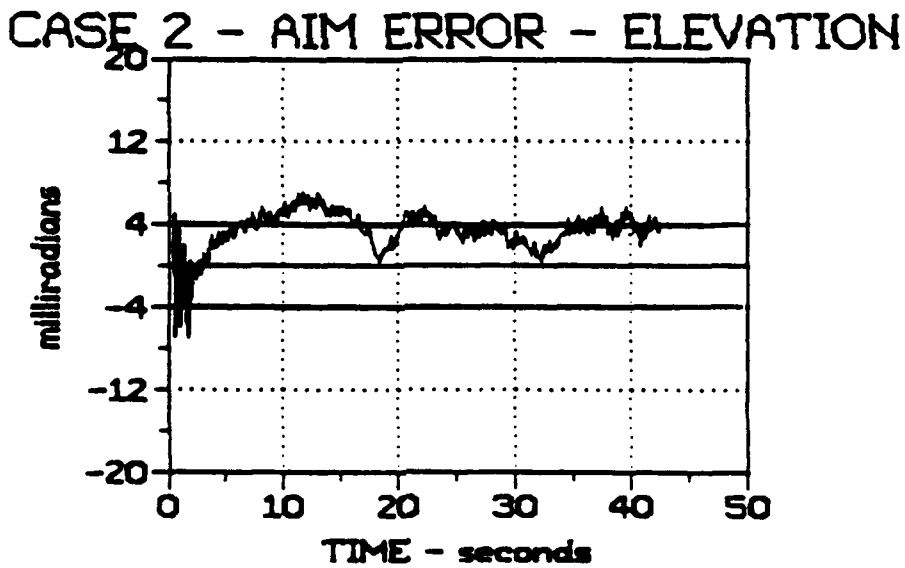
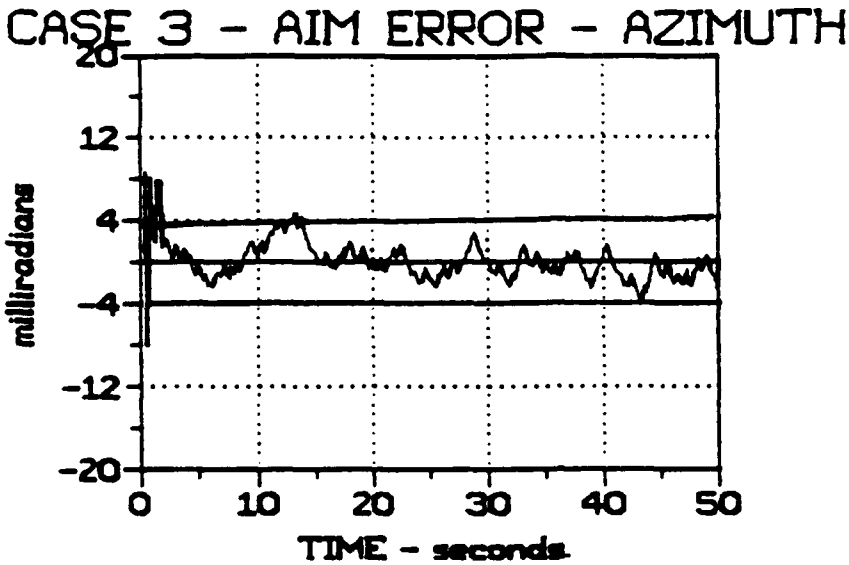


Figure VI-2

Altitude = 250 ft, Velocity = 175 kts, Minimum Range = 2000 mtr



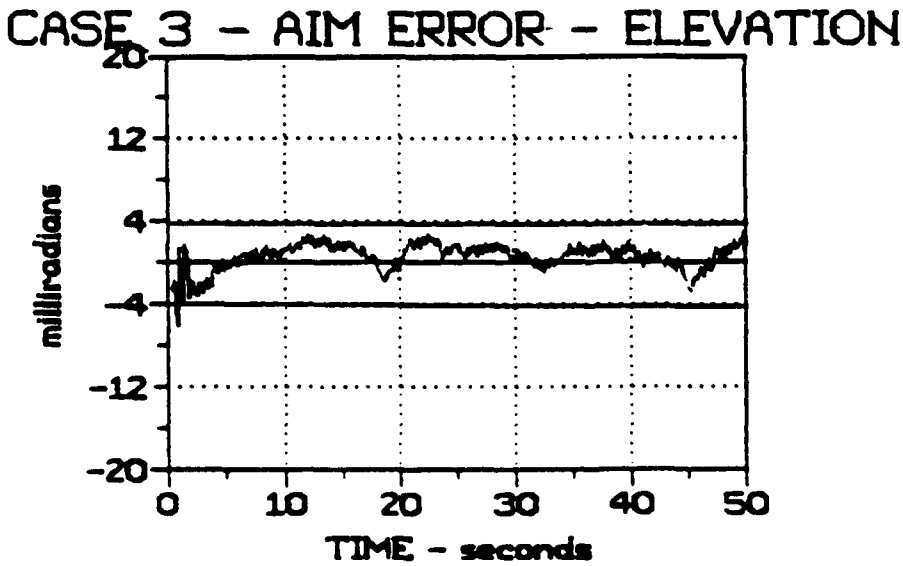
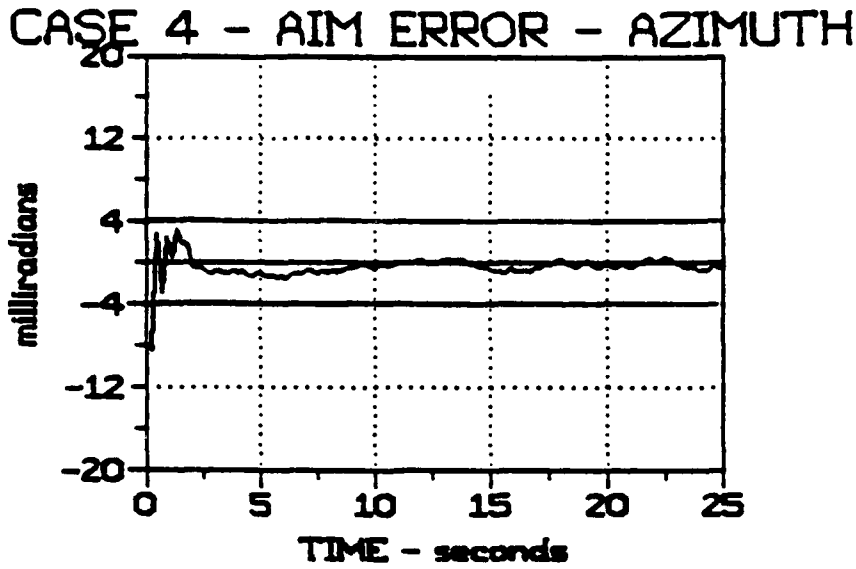


Figure VI-3

Altitude = 250 ft, Velocity = 100 kts, Minimum Range = 1500 mtr



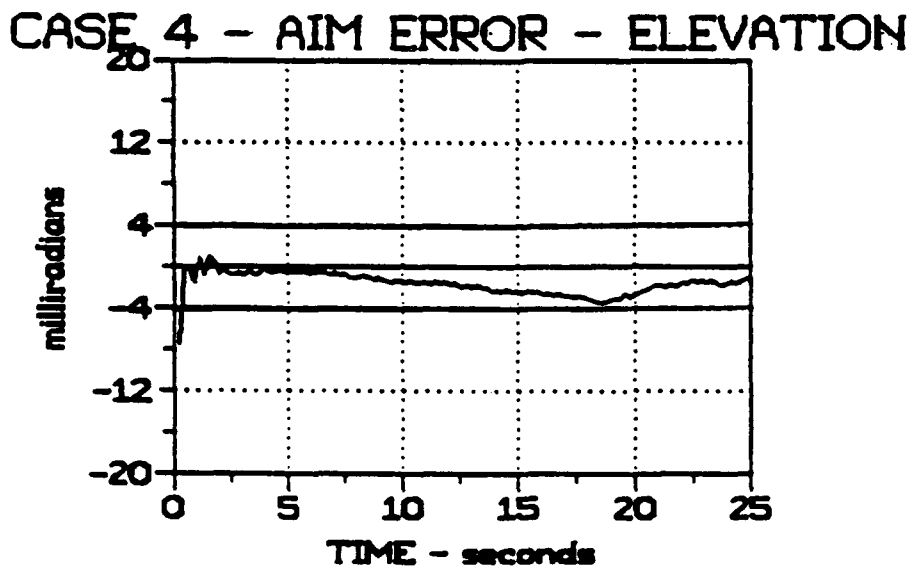
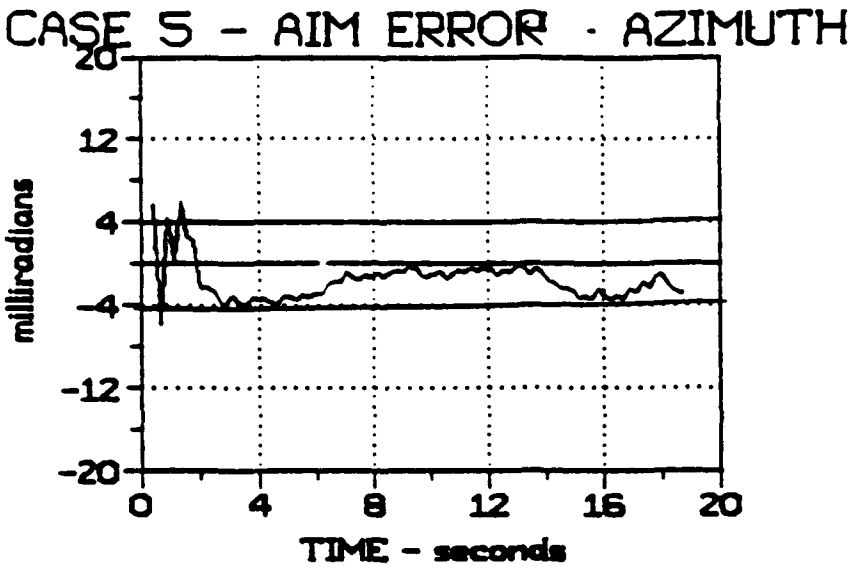


Figure VI-4

Altitude = 250 ft, Velocity = 75 kts, Minimum Range = 500 mtr



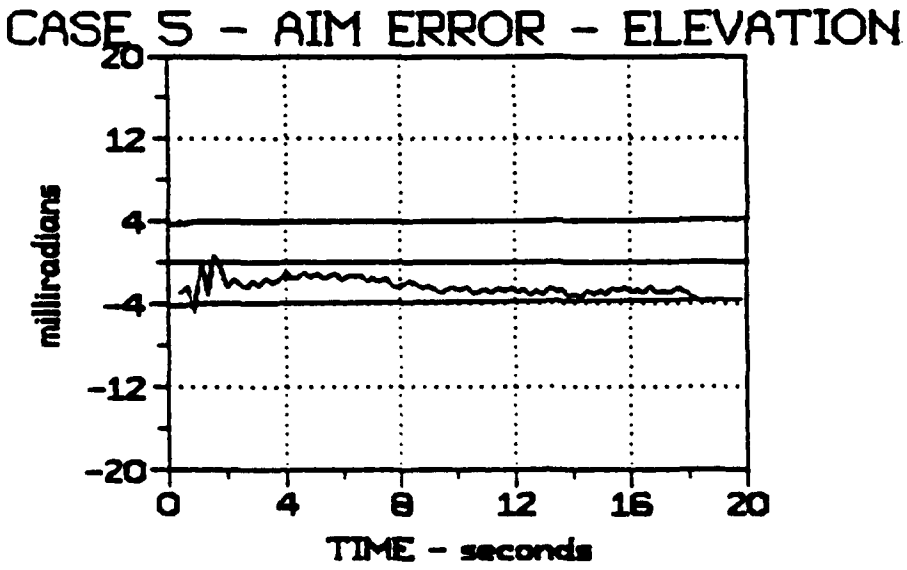


Figure V-5

Altitude = 250 ft, Velocity = 200 kts, Minimum Range = 1000 mtr

CHAPTER VII

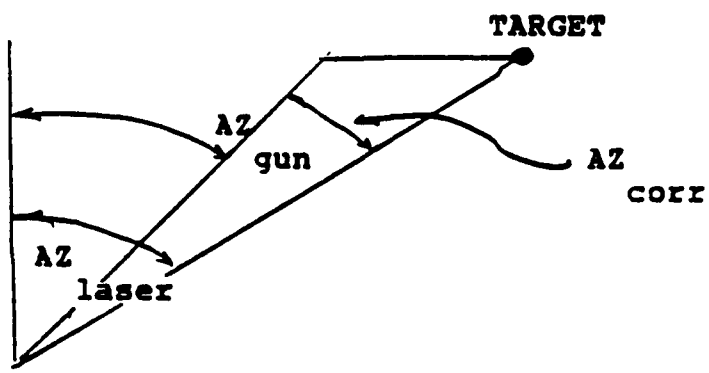
GUN BORESIGHT VERSUS LASER BASED MEASUREMENTS

Theoretical Considerations

A brief analysis was conducted to determine the magnitude of the theoretical error that results when the elevation and azimuth rates are measured along the gun boresight axis versus the laser LOS axis.

AZIMUTH CORRECTION

Referring to the following figure:



If azimuth rate is measured along the laser beam axis

$$AZ_{corr\,los} = - \frac{RSL}{V_{muz}\,Cos(EL_{corr})} \times \dot{AZ}_{los} \qquad \text{(VII-1)}$$

However, as in the BSTING system, when the measurements are made along the gunsight axis

$$AZ_{corr\,gun} = - \frac{RSL}{V_{muz}\,Cos(EL_{corr})} \times \dot{AZ}_{gun} \qquad \text{(VII-2)}$$

The gravity drop correction is the most significant part of the total elevation correction and much larger than elevation rate correction, at least for straight and level flight. Thus it it can be assumed that the Cosine of the elevation correction using laser-based and gun-based measurements are essentially the same. Therefore,

$$AZ_{corr\,error} = - \frac{RSL}{V_{muz}\,Cos(EL_{corr})} \{\dot{AZ}_{los} - \dot{AZ}_{gun}\} \qquad \text{(VII-3)}$$

Now,

$$AZ_{los} = AZ_{gun} + AZ_{corr}$$

so

$$\dot{AZ}_{corr} = \dot{AZ}_{los} - \dot{AZ}_{gun}$$

Taking the differential of equation VII-1

$$\dot{AZ}_{corr} = -\frac{1}{V_{muz}\,Cos(EL_{corr})}\{\dot{RSL} \times \dot{AZ}_{los} + RSL \times \ddot{AZ}_{los}\} \qquad \text{(VII-4)}$$

Where

$$RSL^2 = H^2 + R_{min}^2 + (R_{min} - V_{a/c}\,t)^2$$

$$\dot{RSL} = -V_{a/c}\,(R_{min} - V_{a/c}\,t)/RSL$$

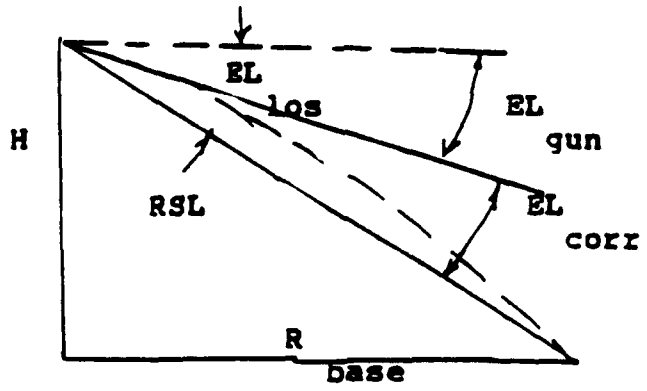
$V_{a/c}$ = Platform Velocity; R_{min} = Range at Closet Approach

$$\dot{AZ}_{los} = V_{a/c}\,R_{min}/RSL^2$$

$$\ddot{AZ}_{los} = -2\,V_{a/c}\,R_{min}\,\dot{RSL}/RSL^3$$

ELEVATION CORRECTION

Referring to the following figure



$$EL_{los} = \text{Arcsin}\{H/RSL\}$$

$$EL_{gd} = f(TOF)$$

$$EL_{gun} = EL_{los} + EL_{gd} \qquad \text{(VII-5)}$$

If measurement of elevation is made along laser axis

$$EL_{corr\ los} = -RSL \times \dot{EL}_{los}/V_{muz} \qquad \text{(VII-6)}$$

However, in the BSTING system

$$EL_{corr\ gun} = -RSL \times \dot{EL}_{gun}/V_{muz} \qquad \text{(VII-6)}$$

Thus the error in the elevation correction becomes

$$EL_{corr\ error} = -RSL\{\dot{EL}_{los} - \dot{EL}_{gun}\}/V_{muz} \qquad \text{(VII-7)}$$

Substituting equation VII-5 into VII-7

$$EL_{corr\ error} = RSL\ \dot{EL}_{gd}/V_{muz} \qquad \text{(VII-8)}$$

The results of these theoretical approximations are shown in figures VII-1 through VII-2. Again the scenarios used are those summarized in Table V-2. The magnitude of these errors are roughly in agreement with the aiming errors resulting from noisy data representing actual measurement made by gunners.

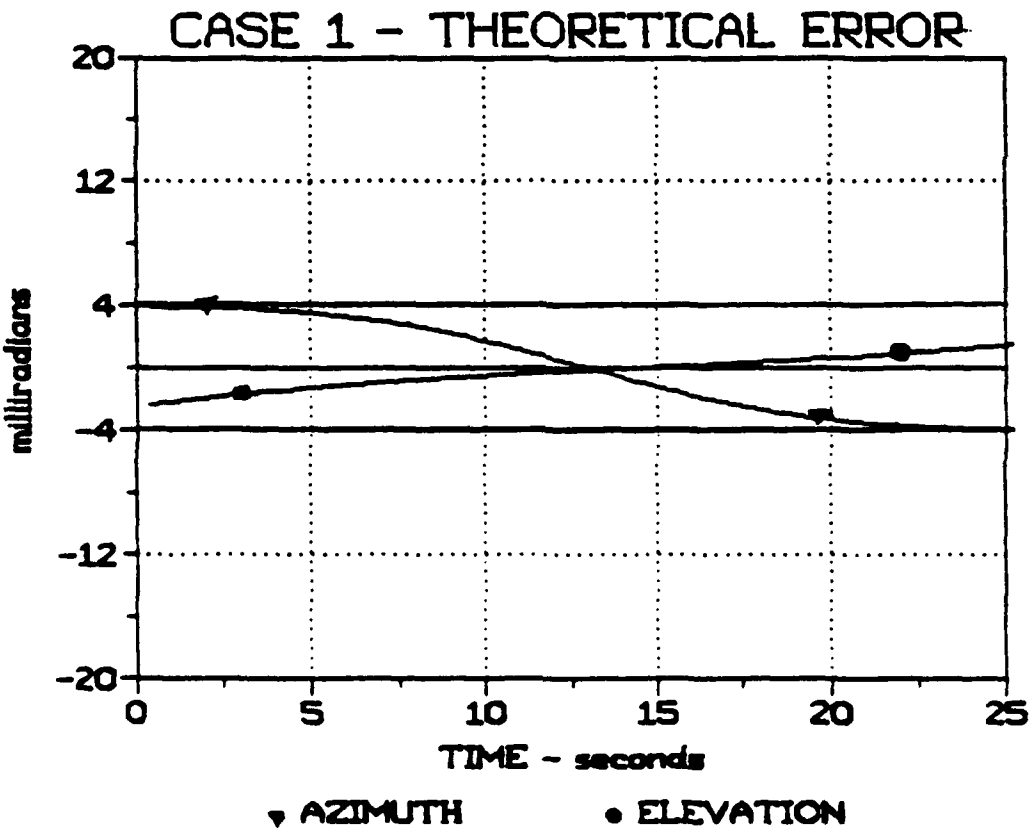


Figure VII-1

Altitude = 250 ft, Velocity = 150 kts, Minimum Range = 1000 mtr

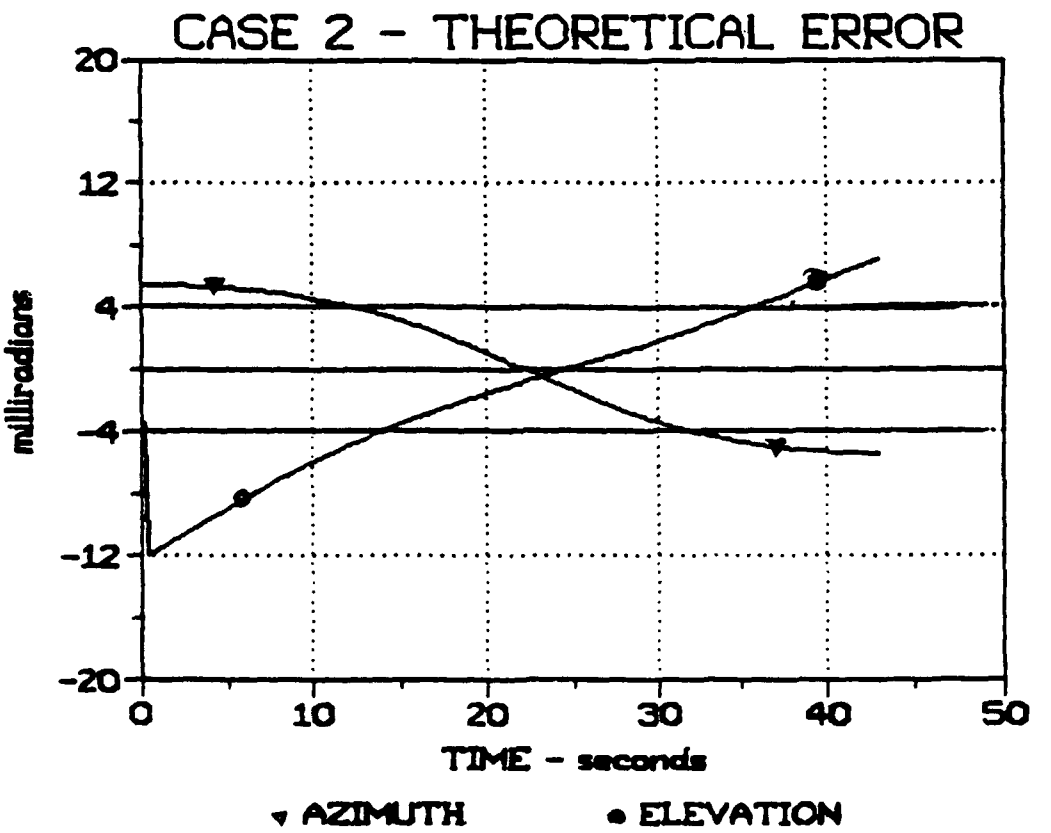


Figure VII-2

Altitude = 250 ft, Velocity = 175 kts, Minimum Range = 2000 mtr

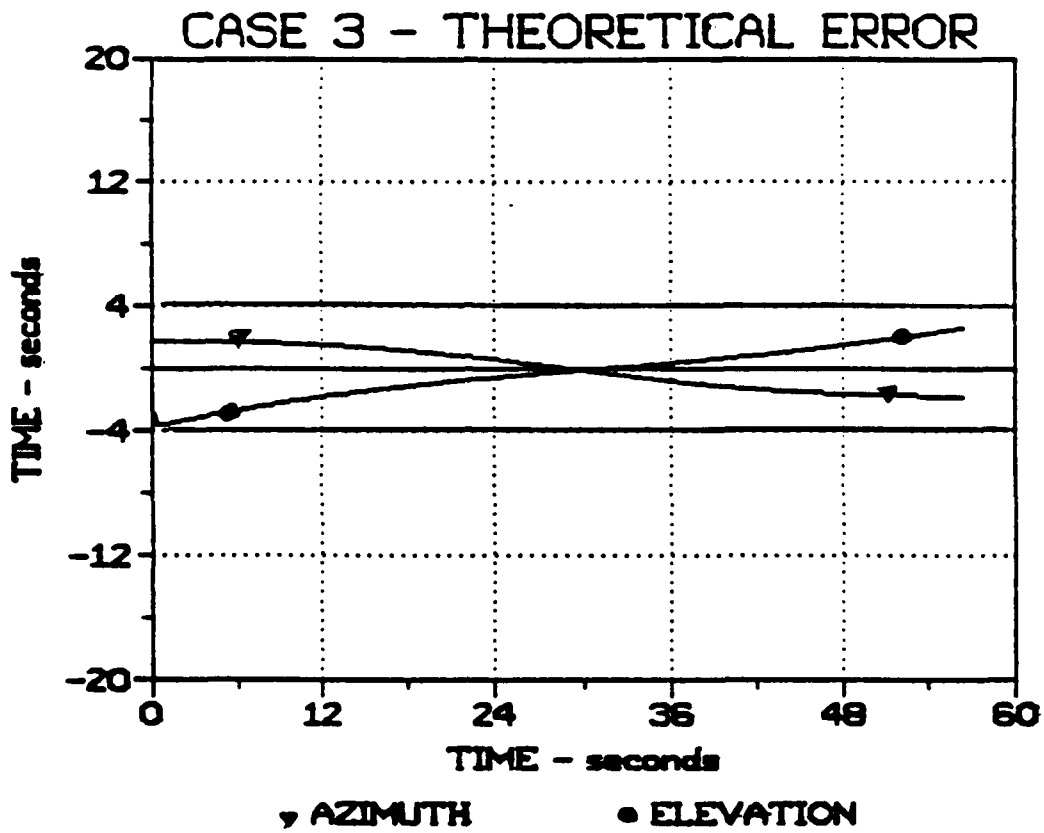


Figure VII-3

Altitude = 250 ft, Velocity = 100 kts, Minimum Range = 1500 mtr

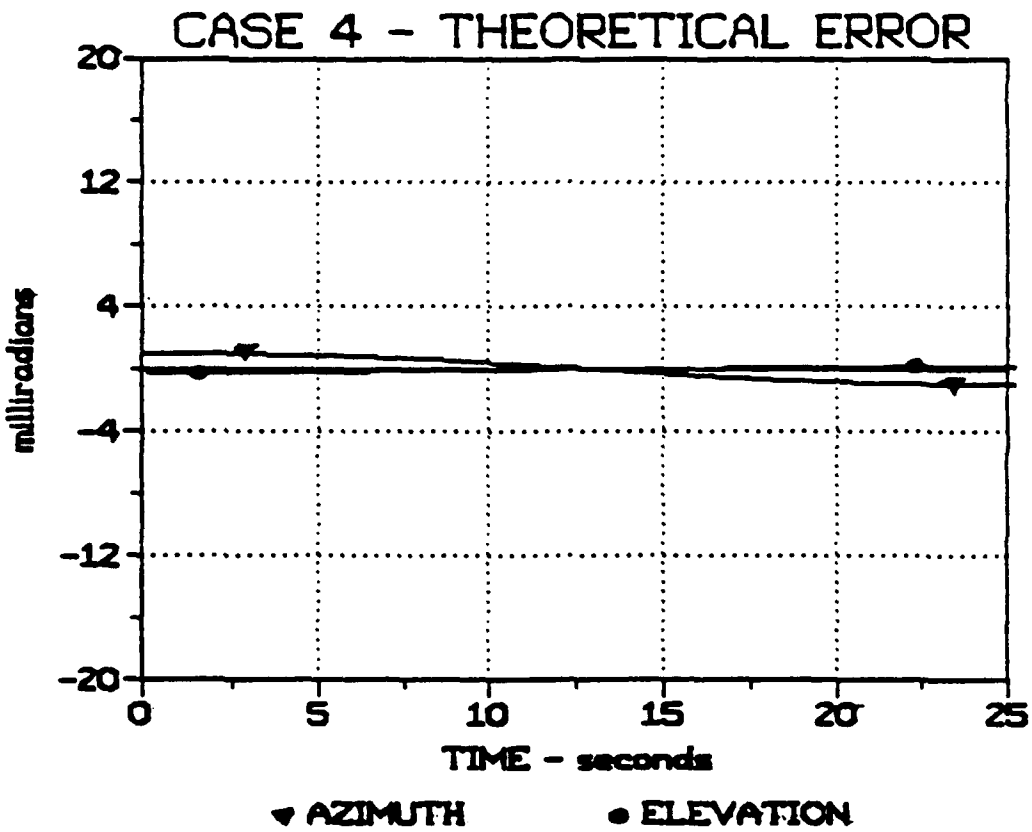


Figure VII-4

Altitude = 250 ft, Velocity = 75 kts, Minimum Range = 500 mtr

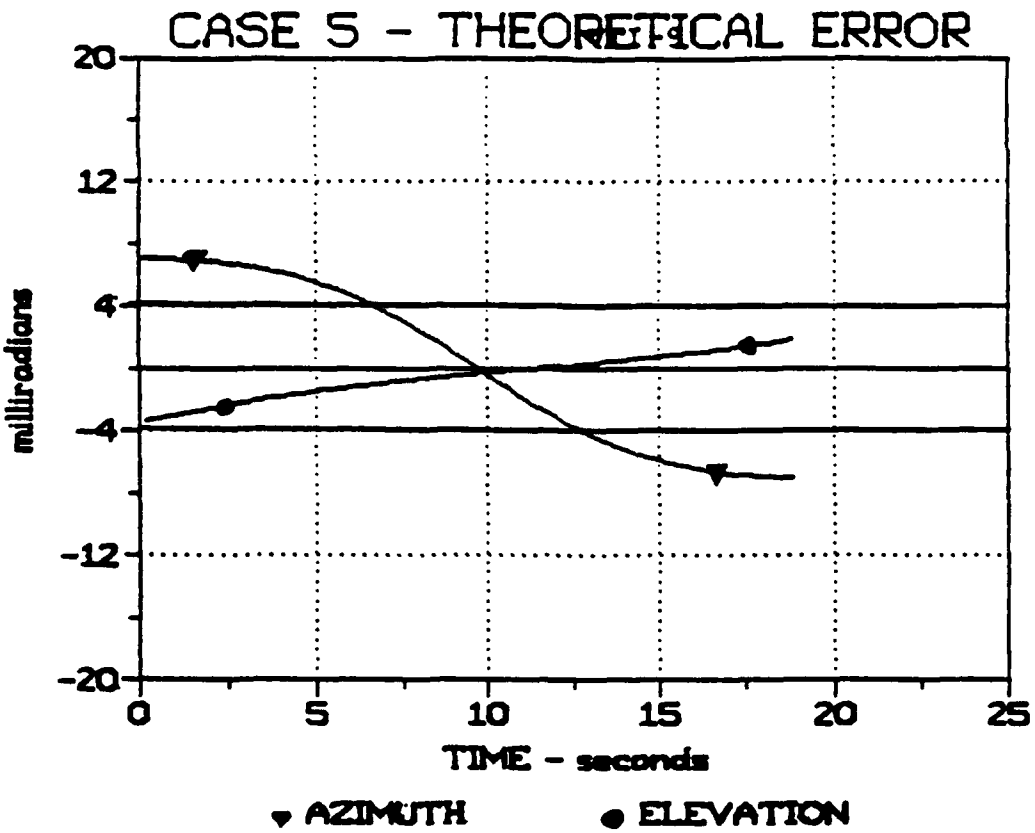


Figure VII-5

Altitude = 250 ft, Velocity = 200 kts, Minimum Range = 1000 mtr

COMPUTER PROGRAM: ERRORPR.BAS

PURPOSE: Computes the Theoretical Azimuth and Elevation Aiming Correction Error when Azimuth/Elevation Rate Measurements are Made along Gun Boresight Axis versus Laser Beam Axis.

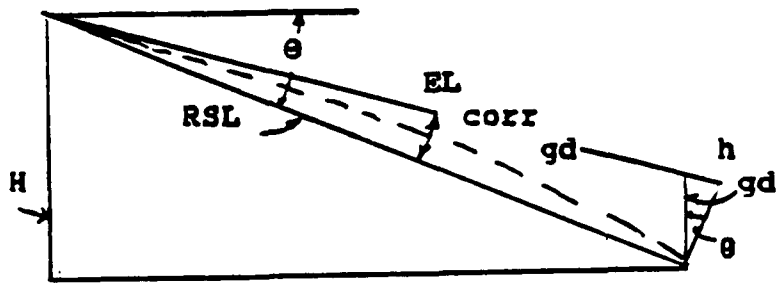
RESULTS: Figures VII-1 to VII-5

```
10 OPEN "O", #1, "B:AZCORRERR.CS4"
20 OPEN "O", #2, "B:ELCORRERR.CS4"
30 PI = 3.1416
40 G = 32.17
45 VMUZ = 2840
50 INPUT "Aircraft Velocity (knots) = ?", VAC
60 VAC = VAC*88/60/.8684
70 INPUT "Minimum Range (meters) = ?", RMIN
80 RMIN = RMIN*3.28
90 INPUT "Altitude (feet) = ?", H
100 RBASE = RMIN - VAC*T
110 RSL = SQR(H^2 + RMIN^2 + RBASE^2)
120 AZ = ATN(RBASE/RMIN)
130 AZ1 = AZ*180/PI
140 EL = ATN(H/SQR(RMIN^2 + RBASE^2))
150 EL1 = EL*180/PI
160 RSLDOT = -VAC*RBASE/RSL
170 AZDOT = VAC*RMIN/RSL/RSL
180 AZDOT2 = -2*VAC*RMIN*RSLDOT/RSL^3
190 ELDOT = -H*RSLDOT/RSL^2
200 ELCORR = -RSL*ELDOT/2840
210 VP = RSLDOT + 2840*COS(ELCORRT)*COS(AZCORR)
220 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
230 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
235 RHO = .076479
240 C = 1.5*RHO/WCDA
250 TOF = RSL/VP/(1 - C*RSL/4)^2
260 ELGD = G*(BETA*TOF)^2/RSL/2
270 ELGDDOT = (ELGD - ELGD1)/.2
280 ELGD1 = ELGD
290 ELCORRT = ELCORR + ELGD
300 AZCORRDOT = -RSLDOT*AZDOT/COS(ELCORRT)/VMUZ - RSL*AZDOT2/VMUZ/COS(ELCORRT)
310 AZCORRERR = -RSL*AZCORRDOT/VMUZ/COS(ELCORRT)
320 ELCORRERR = RSL*ELGDDOT/VMUZ
330 PRINT T, AZ1, EL1, AZCORRERR*1000, ELCORRERR*1000
340 PRINT #1, T, AZCORRERR*1000
350 PRINT #2, T, ELCORRERR*1000
360 IF AZ1 < -43 THEN GOTO 400
370 T = T + .2
380 GOTO 100
400 END
```

CHAPTER VIII

EFFECT OF APPROXIMATIONS IN GRAVITY DROP CORRECTION

The simple BSTING algorithms do not use the elevation angle in gravity correction. Referring to this figure, the gravity drop correction is:



$$h_{gd} = g\, t_f^2 /2$$

Where:

t_f is the time-of-flight

$$EL_{corr_{gd}} = \text{Arctan}\{h_{gd} \text{ Cos } O/RSL\}$$

Two approximations are made in the BSTING algorithms

(1) $\tan EL_{corr_{gd}} = EL_{corr_{gd}}$

(2) $\cos 0 = 1$

Figure VIII-1 shows the elevation angle versus ground range and altitude of the gun platform above the target. The resulting error introduced by these two approximations is shown in figure VIII-2. For altitudes of less than 1000 feet, the error is below 1 milliradian, and in most cases significantly less. If the system is used at higher altitudes, then the elevation angle should be accounted for in the correction. This could be done by either measuring the elevation angle or the actual altitude.

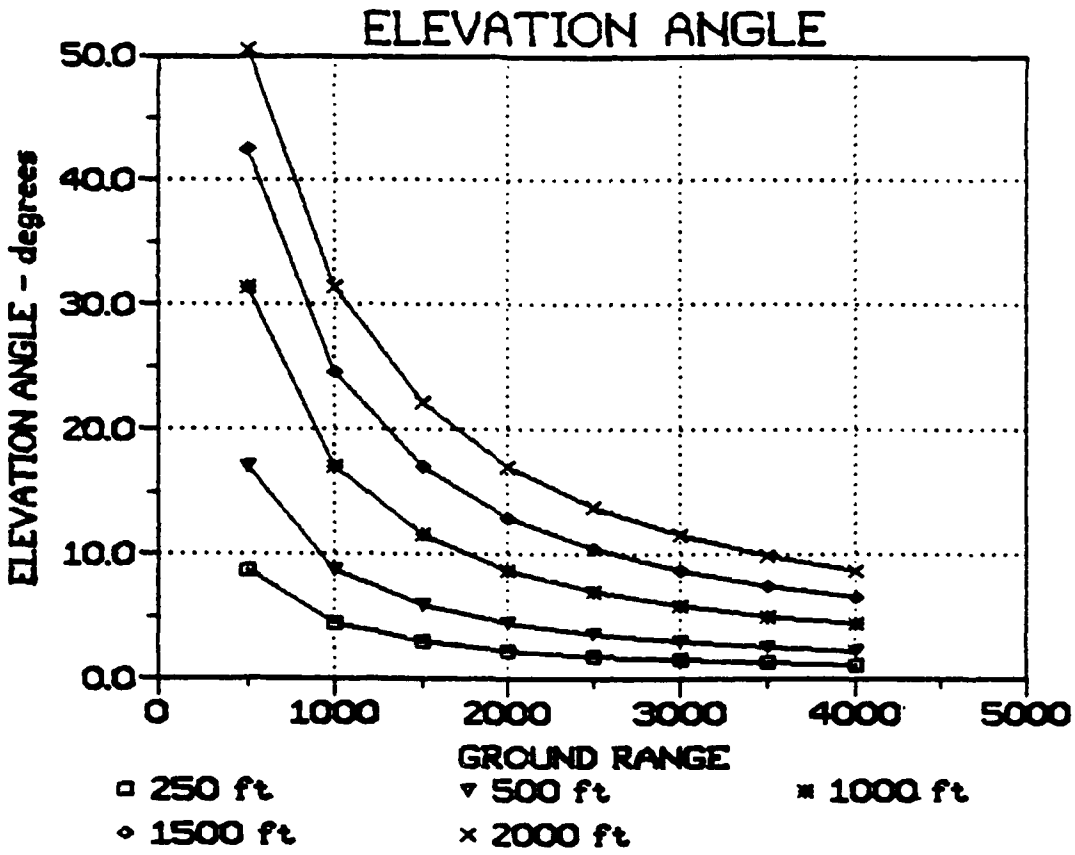


Figure VIII-1

Elevation Angle as a Function of Altitude and Ground Range

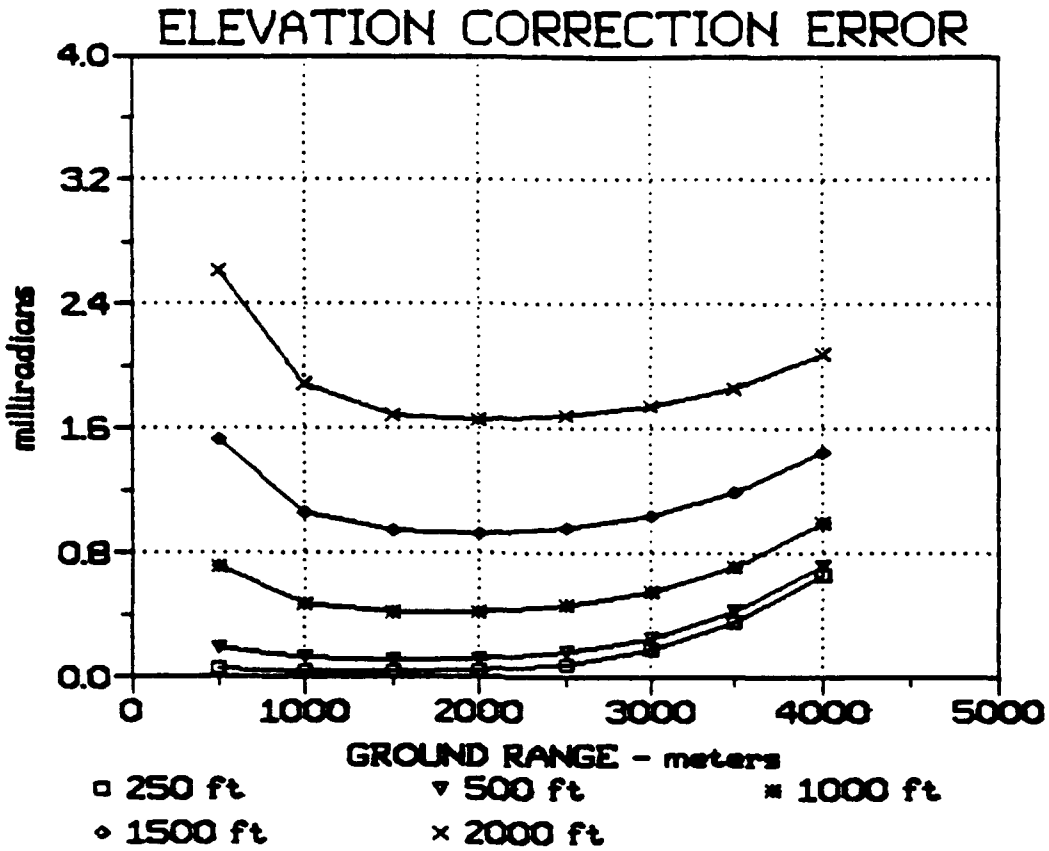


Figure VIII-2

Error in Elevation Aiming Correction Resulting from Neglecting the Actual Elevation Angle (Also the Small Angle Approximation)

COMPUTER PROGRAM: ELERRPR.BAS

PURPOSE: Computes the Elevation Angles as a Function of Altitude and Ground Range. Also Calculates the Error in the Azimuth Aiming Correction Resulting from Neglecting the Elevation Angle (Cos 0 = 1) and by Using the Small Angle Approximation

$$(\tan EL_{corr_{gd}} = EL_{corr_{gd}})$$

RESULTS: Fig. VIII-1 and Fig. VIII-2

```
010 OPEN "O", #1, "C:ELCORRER.100"
20 OPEN "O", #2, "C:EL.100"
30 G = 32.17
40 PI = 3.1416
50 VMUZ = 2840
60 INPUT "Minimum Ground Range (meters) = ?", RMIN
70 INPUT "Altitude (feet) = ?", H
80 RBASE = RMIN
90 RSL = SQR(H^2 + (RBASE*3.28)^2)
100 EL = ATN(H/RBASE/3.28)
110 VP = VMUZ
120 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
130 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
135 RHO = .076479
140 C = 1.5*RHO/WCDA
150 TOF = RSL/VP/(1 - C*RSL/4)^2
160 ELGD = G*(BETA*TOF)^2/RSL/2
170 ELGD1 = ATN(ELGD*COS(EL))
180 DIFF = ELGD - ELGD1
190 PRINT RBASE, TOF, ELGD*1000, ELGD1*1000, DIFF*1000, EL*180/PI
200 PRINT #1, RBASE, DIFF*1000
205 PRINT #2, RBASE, EL*180/PI
210 IF RBASE > 3750 THEN GOTO 240
220 RBASE = RBASE + 500
230 GOTO 90
240 END
```

CHAPTER IX

FIRE CONTROL SYSTEM USING EXTERNAL MEASUREMENTS

Theoretical Design

An alternate design of a fire control system would use external measurement of parameters in lieu of (or in addition to) sensors mounted on the gun itself as is the case with the current BSTING system. The goal is a more responsive fire control system that can provide "point and shot" capability.

SYSTEM CONCEPT

This initial system design uses the following sensors for measuring the parameters needed to make the required azimuth and elevation aiming corrections.

a. The slant range would be measured by a laser rangefinder.

b. Angular resolvers mounted on the gun mount would sense the azimuth and elevation of the gun with respect to the platform. As in the BSTING system, angular measurements are made along the gun boresight axis rather than the more desirable laser beam axis.

c. An inertial platform would also be needed to reference the platform with respect to inertial space in the vertical and horizontal directions.

d. Platform velocity would be obtained from the aircraft navigation system or externally from the Global Positioning Satellite (GPS) system.

Again the key source of error in this system, as with the BSTING system, is that the azimuth and elevation data has to be measured with respect to the gun boresight rather than the laser beam axis. An analysis was conducted to determine the magnitude of this error.

<u>LOS Range</u>. In this FCS the laser range finder would be measuring the slant range to the target. For straight and level flight, the range measured would be based on the relationship:

$$RSL^2 = H^2 + RMIN^2 + (RMIN - VAC \times t)^2 \qquad \text{(IX-1)}$$

Where

H = Altitude above target

RMIN = Ground range at point of closest approach

VAC = Aircraft velocity

t = Time

LOS Elevation and Azimuth Angles. The elevation angle is determined from (refer to figure IX-1a)

$$EL_{los} = \text{Arcsin}\{H/RSL\} \qquad \text{(IX-2)}$$

and the azimuth angles is (Refer to figure IX-1b)

$$AZ_{los} = \text{Arctan}\{(RMIN - VAC \times t)/RMIN\} \qquad \text{(IX-3)}$$

Incidently, unlike the BSTING system where the azimuth plane is defined along the gun boresight azimuth, in this system both the elevation and azimuth planes are defined in inertial space.

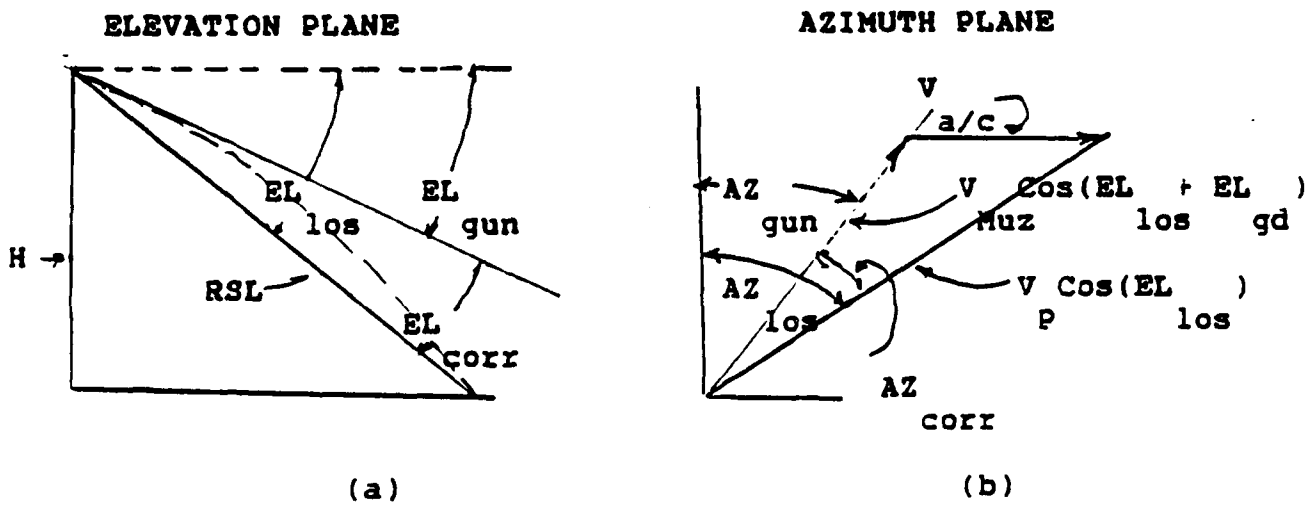


Figure IX-1

Geometry For Theoretical Analysis

Elevation Aiming Correction. In the elevation plane, the correction was determined from:

$$El_{corr} = 0.5\ G \times Cos(EL_{los}) \times (\ \times T_f^2)\ /RSL \qquad \text{(IX-4)}$$

Where:

$$T_f = \frac{RSL}{V_p\ (1 - c \times RSL/4)^2}$$

$$c = 1.5\ Density/WCDA$$

$$WCDA = .000005 \times RSL^2 - 0.036938 \times RSL + 700.4299$$

$$= 3E\text{-}09 \times RSL^2 - 4.706E\text{-}05 \times RSL + 1.0216$$

$$Density = 0.07479$$

Azimuth Aiming Correction. Referring to figure IX-1b, the required corrections in the azimuth plane are determined by:

$$Sin(AZ_{gun}) = \frac{V_p\ Sin(AZ_{las})Cos(EL_{las}) - V_{a/c}}{V_{muz}\ Cos(EL_{las} + EL_{gd})} \qquad \text{(IX-5)}$$

and

$$AZ_{corr} = AZ_{las} - AZ_{gun}$$

$$V_p^2 = V_{muz}^2 + 2\ V_{muz}\ V_{a/c}\ Sin(AZ_{las})Cos(EL_{las}) + V_{a/c}^2$$

In this initial investigation, the inertial platform was not included as it was assumed that the aircraft was flying parallel to the earth. The scenarios shown in Table IX-I were used to test this system.

TABLE IX-I

SCENARIOS

CASE	ALTITUDE (feet)	VELOCITY (knots)	MINIMUM RANGE (meters)
1	250	150	1000
2	250	175	2000
3	250	100	1500
4	250	75	500
5	250	200	1000

These scenarios are the same ones used in evaluating the BSTING system, as reported in the previous chapters. The results are shown in figures IX-3 through IX-12. The curves marked "Laser" represent the desired correction, in other words the correction that theoretically provided the greatest accuracy. The curves marked "GUN" represent the actual corrections that result when azimuth and elevation angles are measured along the boresight axis. The difference

between these two values represents the "UNCORRECTED" error in aiming results when the boresight-based azimuth and elevation angles measured were used to compute corrections. Reviewing these figures, the maximum aiming errors are shown in Table IX-II.

TABLE IX-II

MAXIMUM AIMING ERRORS (UNCORRECTED)

SCENARIO	AZIMUTH (milliradians)	ELEVATION (milliradians)
1	-3.5	-0.2
2	-4.5	-0.8
3	-1.4	-0.15
4	-1.0	0.0
5	+/-6.0	-0.35

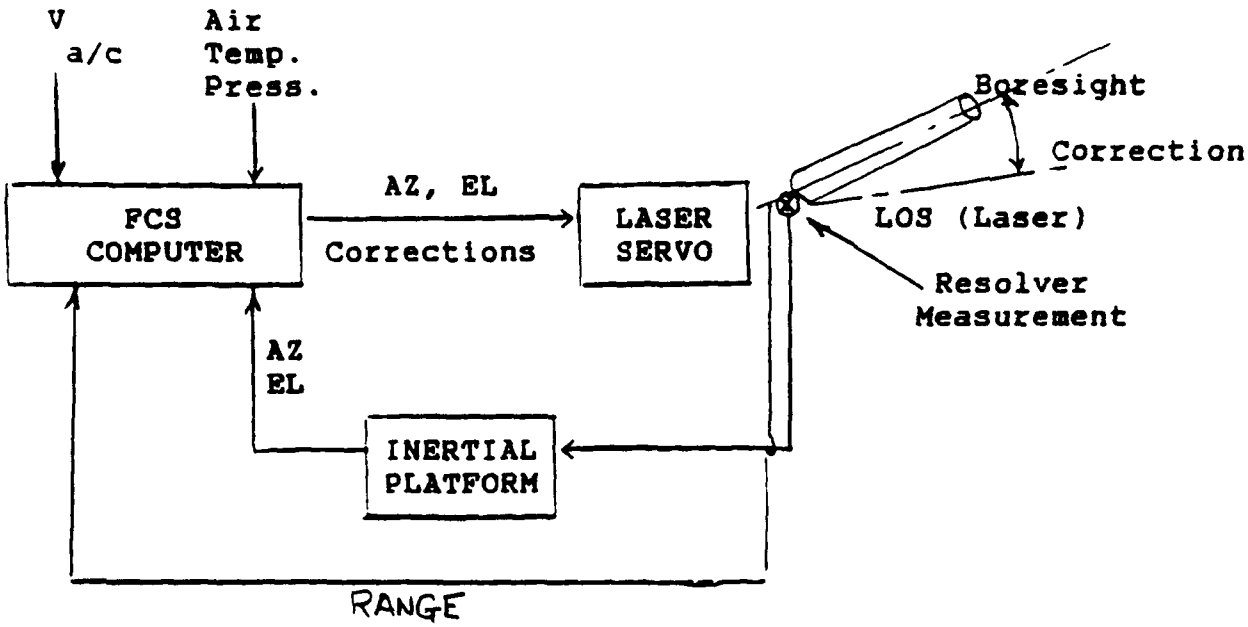
It can be seen from this data that while errors in elevation are minimal, they are significant in azimuth. Therefore, a scheme is required to correct the azimuth aiming corrections. As an initial technique, azimuth/elevation measurements along the gun boresight axis were corrected by the value of the previous correction before they were used to compute the next correction, i.e.

$$AZGUN_{k+1} = AZGUN_{k} + AZCORR_{k}$$

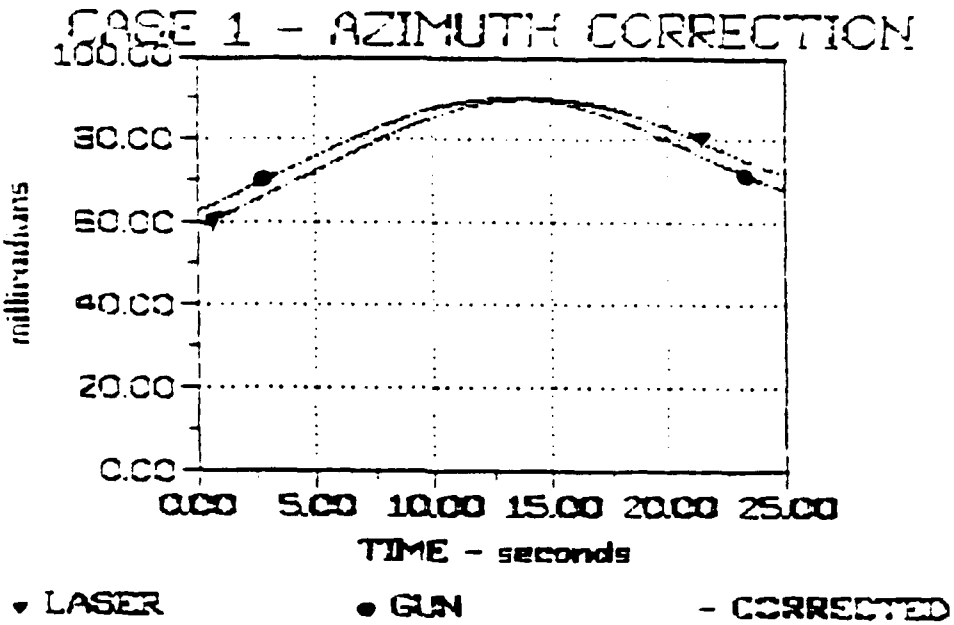
and

$$ELGUN_{k+1} = ELGUN_{k} + ELCORR_{k}$$

Figure IX-2 is a block schematic diagram showing how this technique might be implemented.



The results of implementing this correction technique is shown in figure IX-3 through IX-11 and identified as "CORRECTED" corrections and aiming error. As can be seen, this technique corrects the error at least for the theoretical case without noisy data. Next, the technique will be investigated in a simulated gunner operated system where the measurements are noisy.



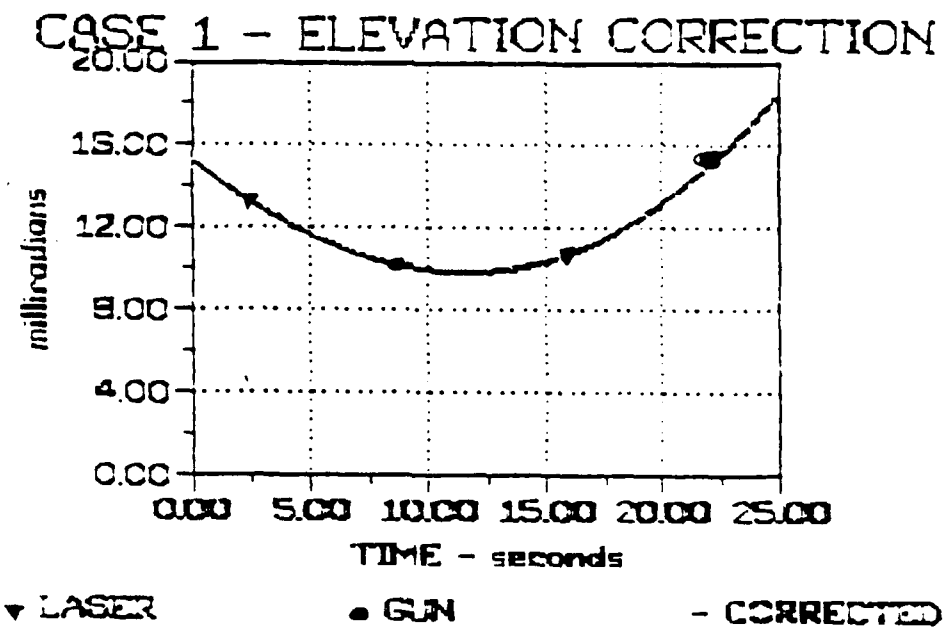
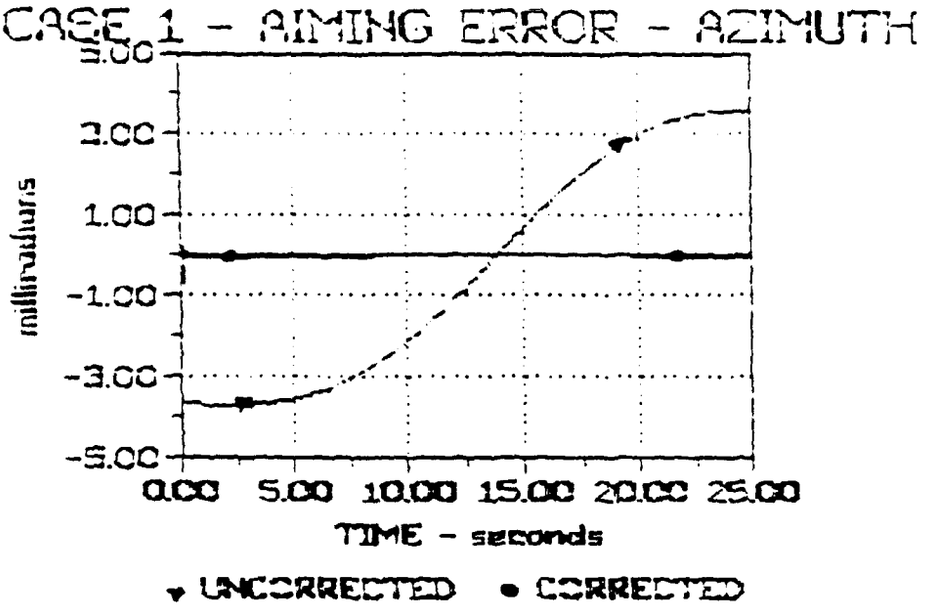


Figure IX-3

Altitude = 250 ft; Aircraft Velocity = 150 kts;
Minimum Range = 1000 mtr



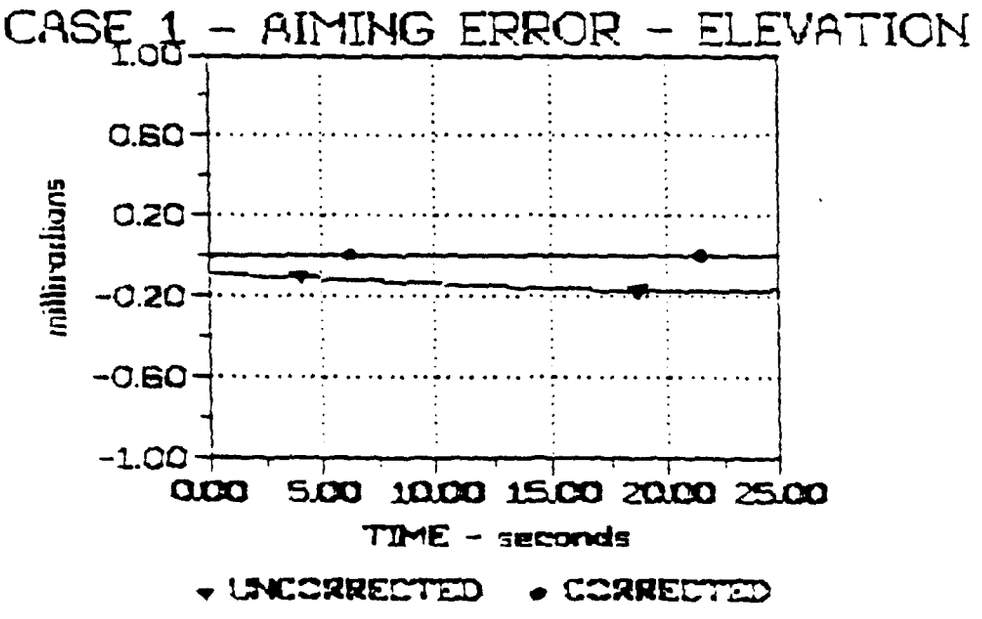
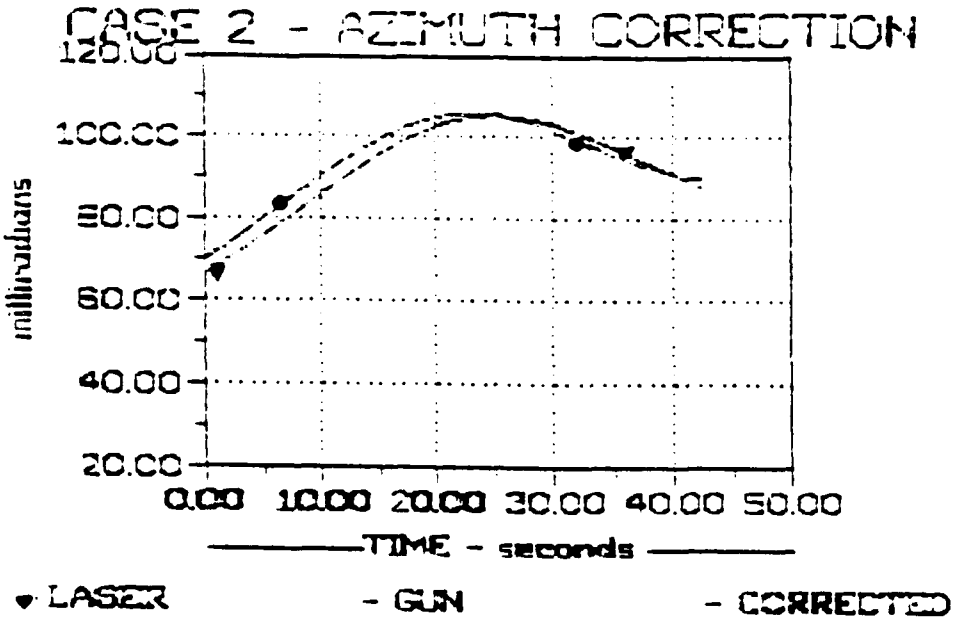


Figure IX-4

Altitude = 250 ft; Aircraft Velocity = 150 kts;
Minimum Range = 1000 mtr



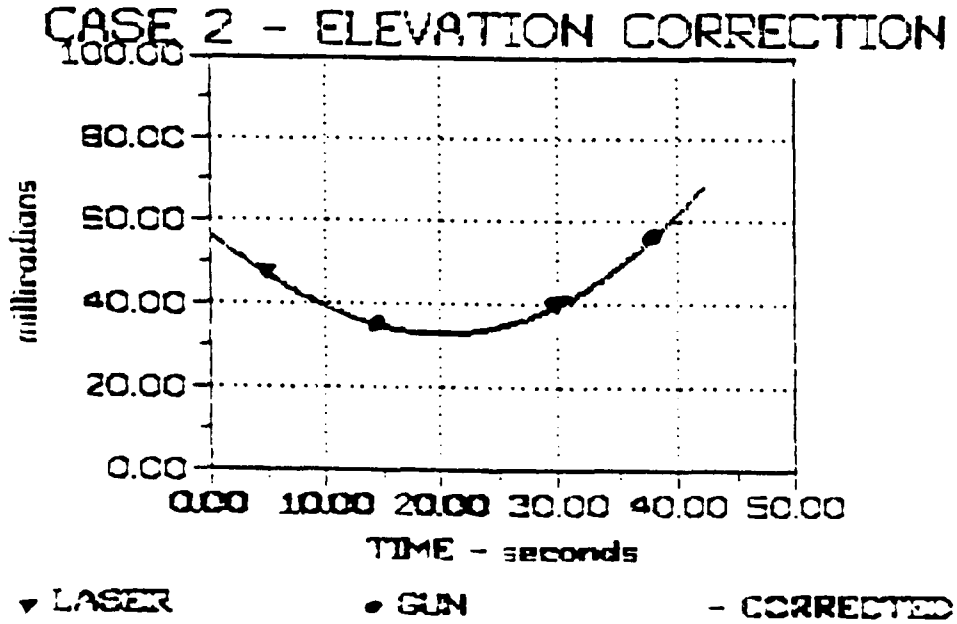
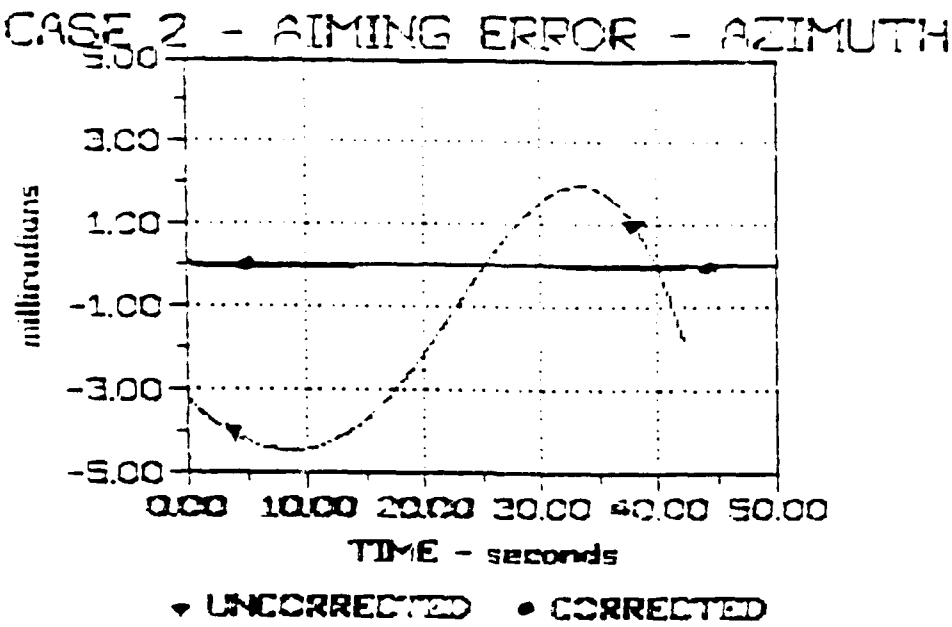


Figure IX-5

Altitude = 250 ft; Aircraft Velocity = 175 kts;
Minimum Range = 2000 mtr



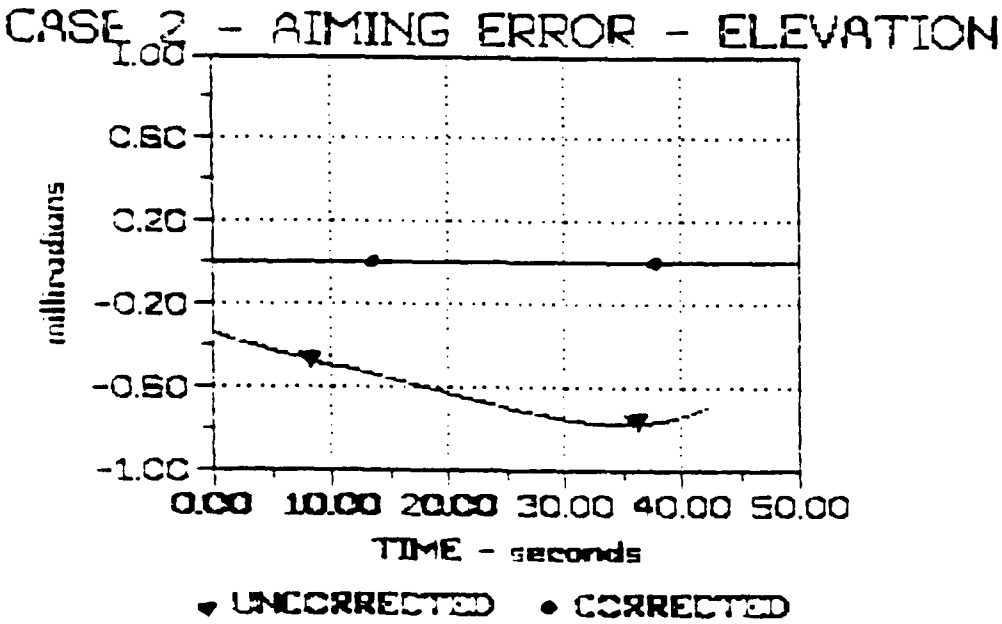
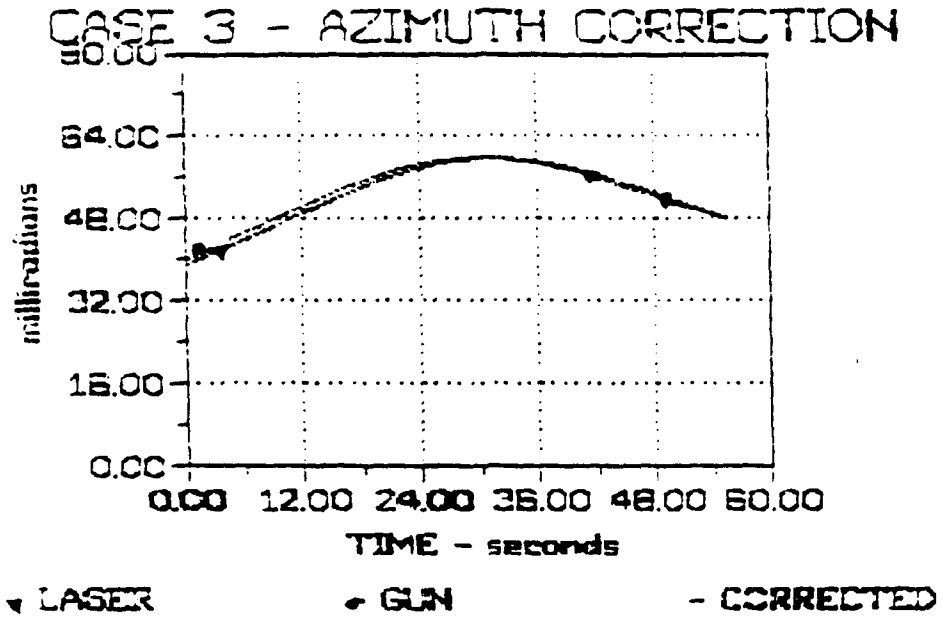


Figure IX-6

Altitude = 250 ft; Aircraft Velocity = 175 kts;
Minimum Range = 2000 mtr



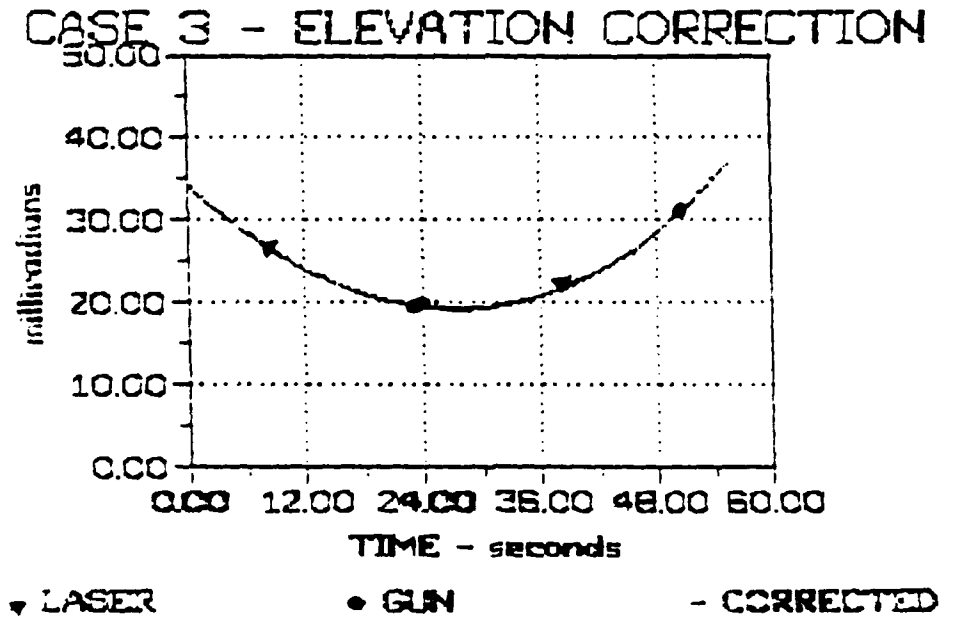
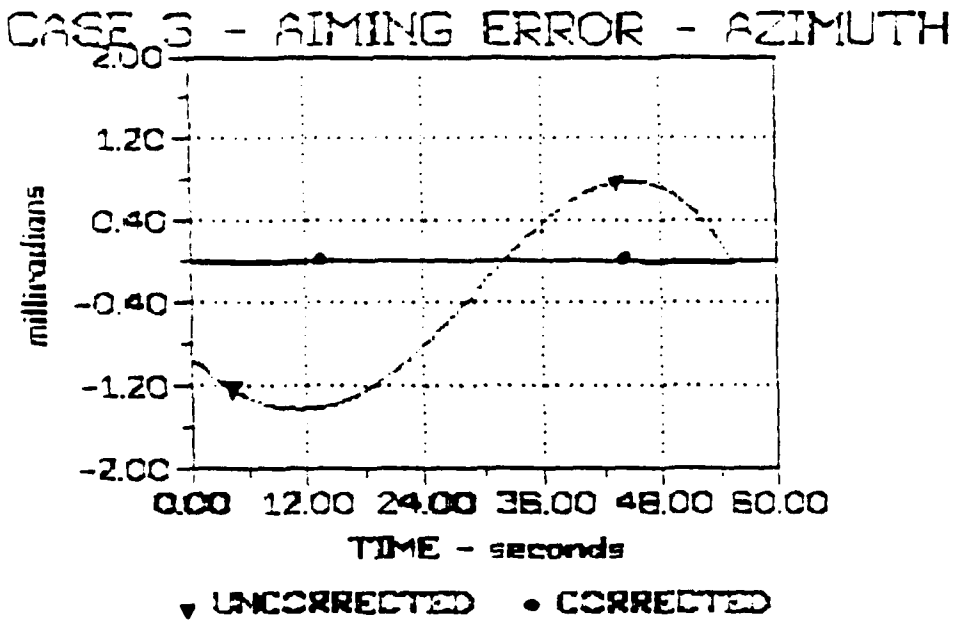


Figure IX-7

Altitude = 250 ft; Aircraft Velocity = 100 kts;
Minimum Range = 1500 mtr



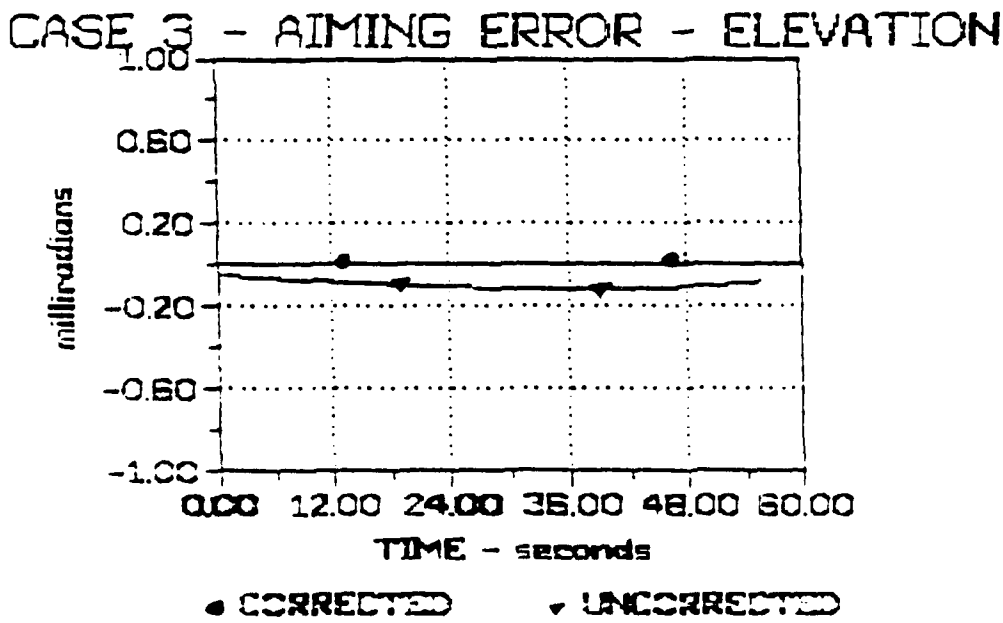
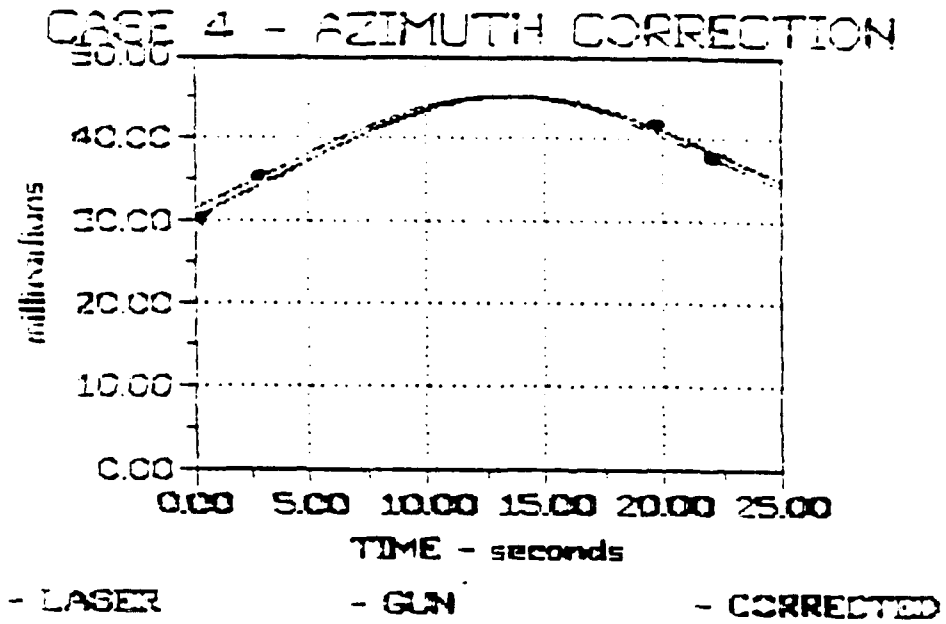


Figure IX-8

Altitude = 250 ft; Aircraft Velocity = 100 kts;
Minimum Range = 1500 mtr



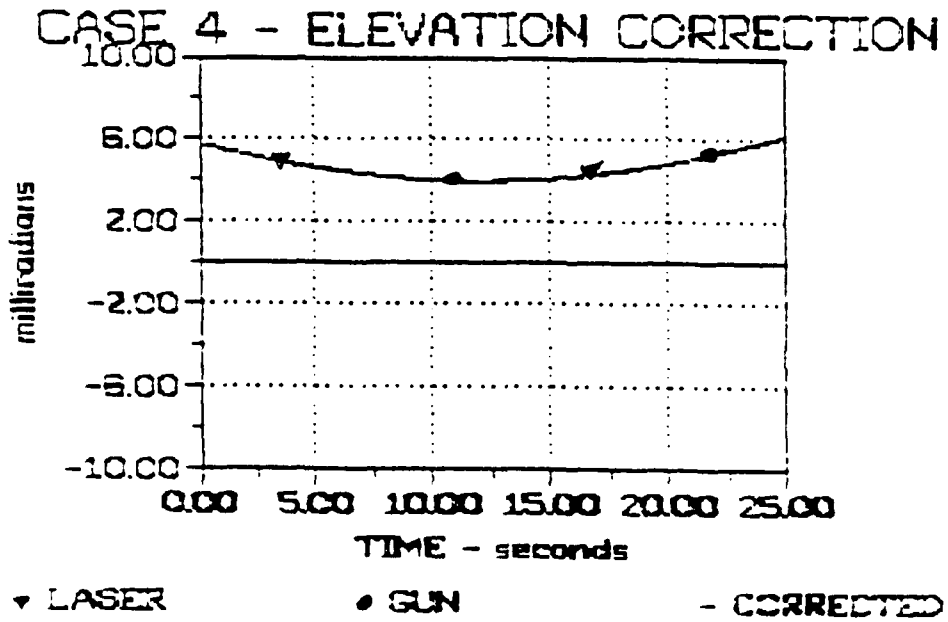
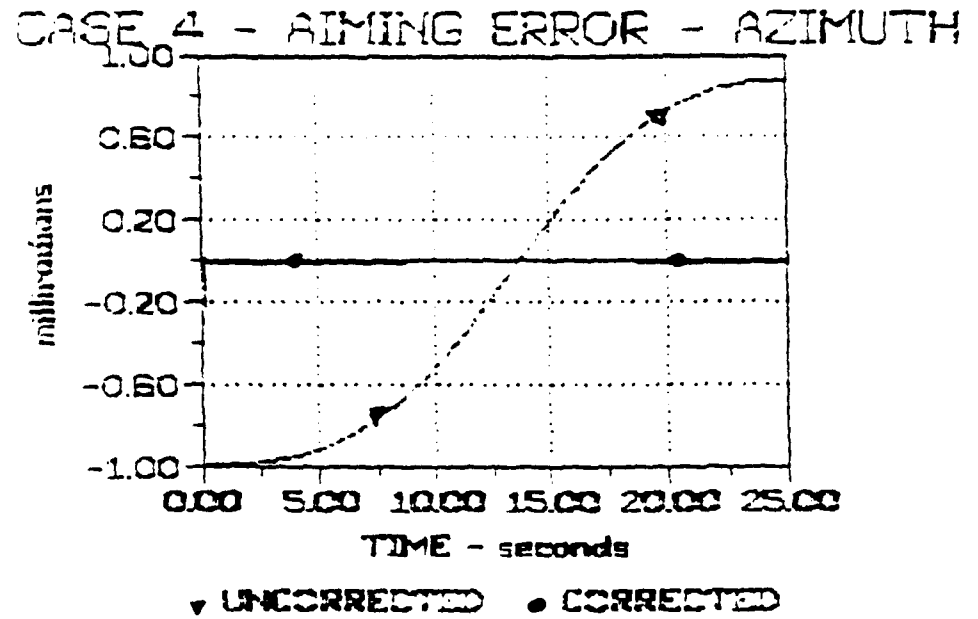


Figure IX-9

Altitude = 250 ft; Aircraft Velocity = 75 kts;
Minimum Range = 500 mtr



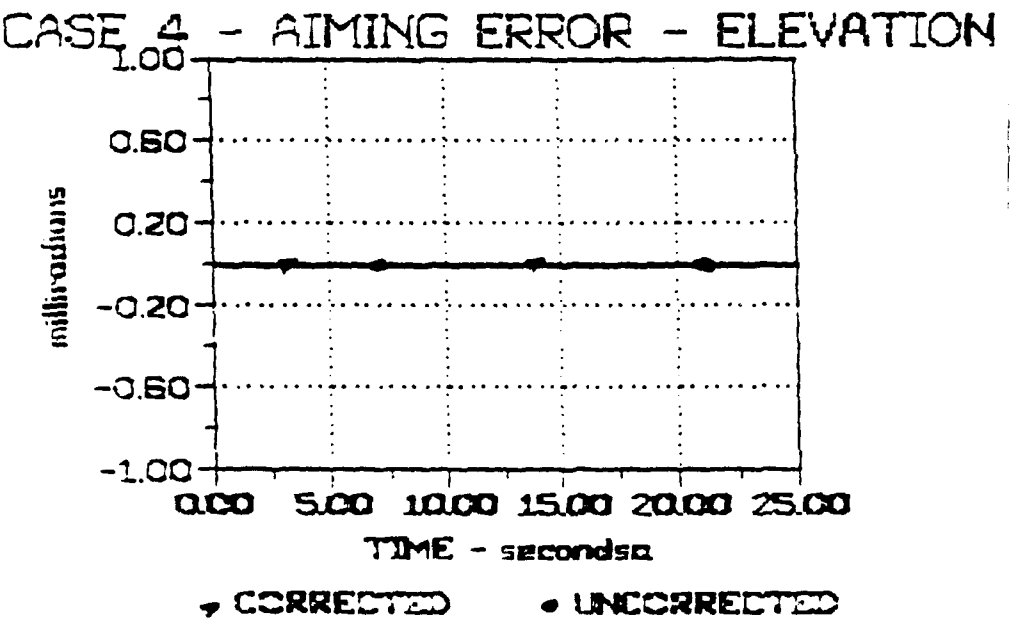
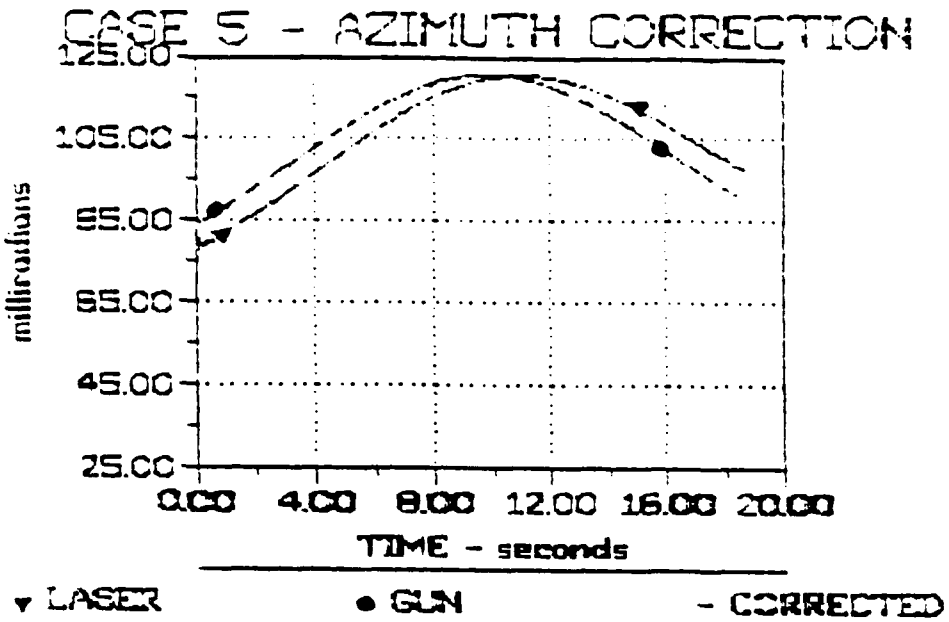


Figure IX-10

Altitude = 250 ft; Aircraft Velocity = 75 kts;
Minimum Range = 500 mtr



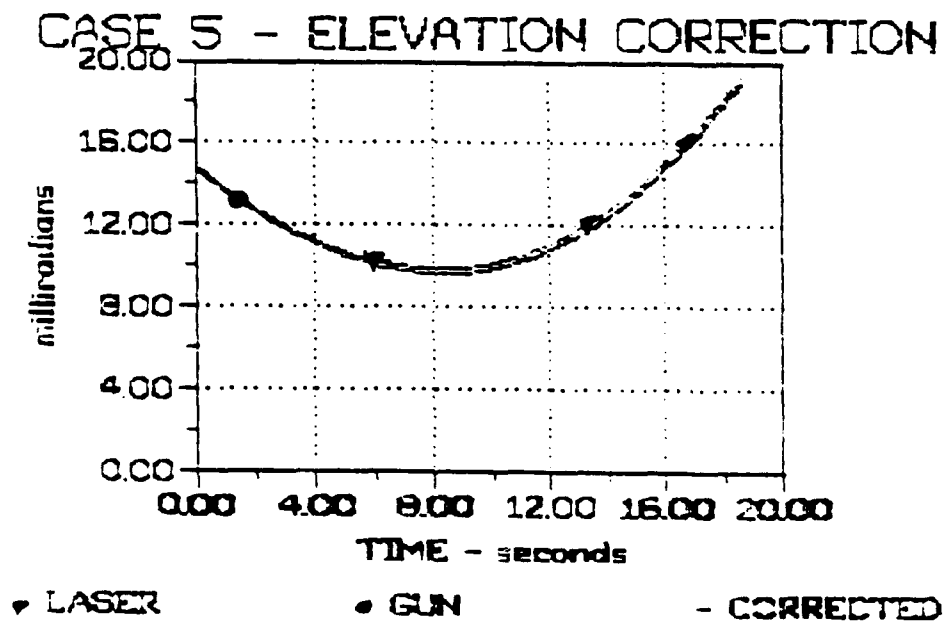
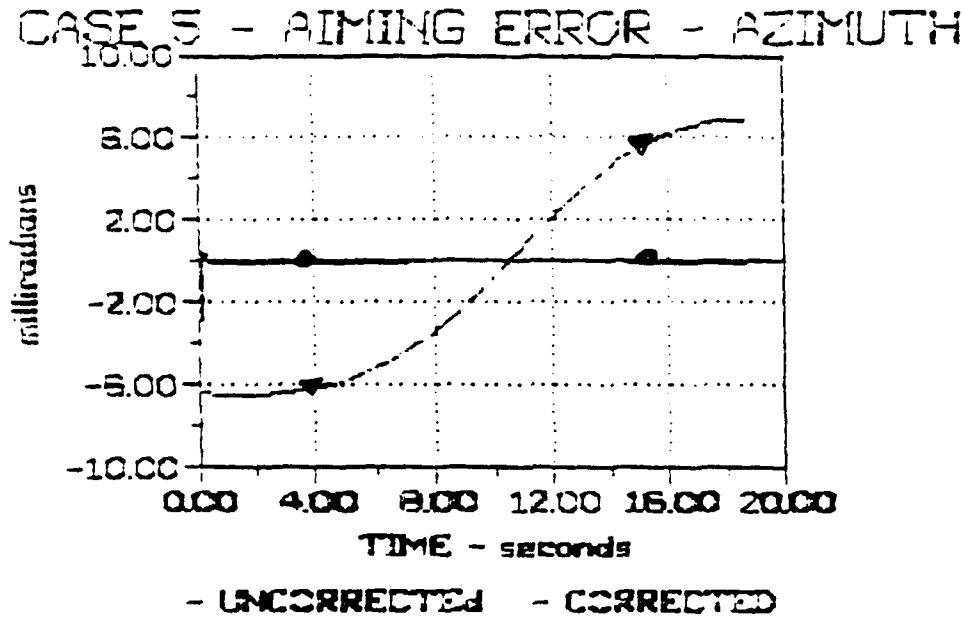


Figure IX-11

Altitude = 250 ft; Aircraft Velocity = 200 kts;
Minimum Range = 1000 mtr



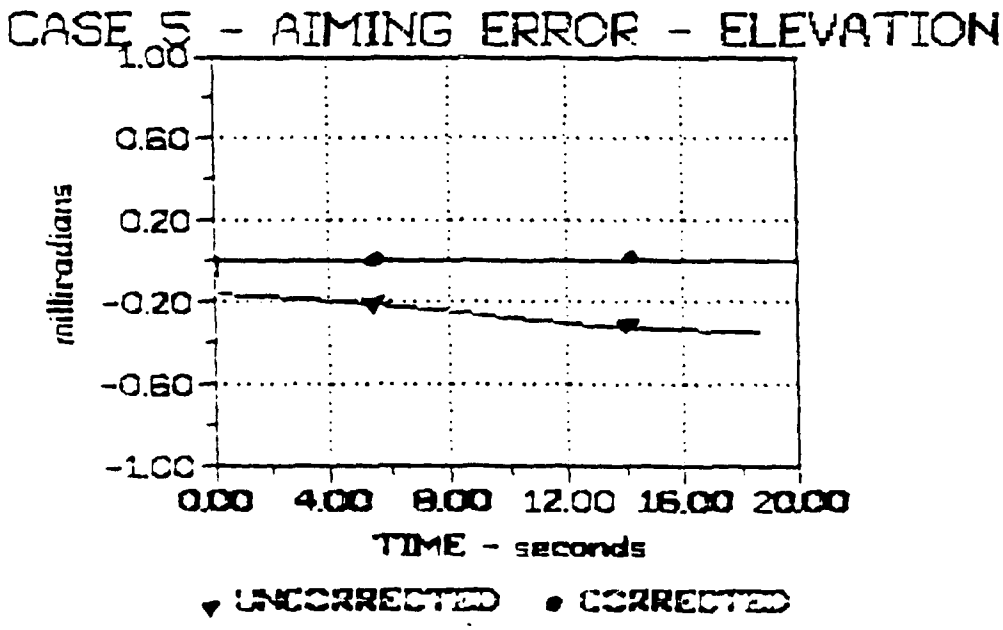


Figure IX-12

Altitude = 250 ft; Aircraft Velocity = 200 kts;
Minimum Range = 1000 mtr

COMPUTER PROGRAM: LOSPR.BAS

PURPOSE: Computes the Line-of-Sight Azimuth/Elevation Angle and Range of the Target with Respect to Platform. These Parameters are "Measured" along Laser Axis.

RESULTS: Used in LASTHEPR.BAS and GUNDATPR.BAS

```
 10 OPEN "O", #1, "B:LOSDATA.CS5"
20 PI = 3.1416
30 DT = .2
40 G = 32.17
50 INPUT "Altitude (feet) = ?", H
60 INPUT "Aircraft Velocity (knots) = ?", VAC
70 VAC = VAC*88/60/.8684
80 INPUT "Minimum Ground Range to Target = ?", RMIN
90 RMIN = RMIN*3.28
100 RBASE = RMIN - VAC*T
110 RSL = SQR(H^2 + RMIN^2 + RBASE^2)
120 AZLOS = ATN(RBASE/RMIN)
130 ELLOS = ATN(H/SQR(RBASE^2 + RMIN^2))
140 PRINT #1, T, AZLOS, ELLOS, RSL
142 PRINT  T, AZLOS*180/PI, ELLOS*180/PI, RSL
150 IF AZLOS*180/PI < -42 THEN GOTO 180
160 T = T + DT
170 GOTO 100
180 END
```

COMPUTER PROGRAM: LASTHEPR.BAS

PURPOSE: Computes the Azimuth and Elevation Corrections Based on Measurements Made along Laser Beam Axis.

RESULTS: Used in Figures IX-2 through IX-11

```
010 OPEN "I", #1, "B:LOSDATA.CS5"
20 OPEN "O", #2, "B:LASAZCOR.CS5"
30 OPEN "O", #3, "B:LASELCOR.CS5"
40 OPEN "O", #4, "B:OUTLAS.CS5"
50 PI = 3.1417
60 G = 32.17
70 VMUZ = 2840
80 INPUT "Aircraft Velocity = ?", VAC
90 VAC = VAC*88/60/.8684
100 INPUT #1, T, AZLAS, ELLAS, RSL
110 VP = SQR(VMUZ^2 + 2*VMUZ*VAC*SIN(AZLAS)*COS(ELLAS) + VAC^2)
120 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
130 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
140 RHO = .07479
150 C = 1.5*RHO/WCDA
160 TOF = RSL/VP/(1 - C*RSL/4)^2
170 ELGD = 16.085*COS(ELLAS)*(BETA*TOF)^2/RSL
180 ELCORR = ELGD
190 AZGUN = VP*SIN(AZLAS)*COS(ELLAS) - VAC
200 AZGUN = AZGUN/VMUZ/COS(ELGD + ELLAS)
210 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
220 AZCORR = AZLAS - AZGUN
230 PRINT T, AZLAS*180/PI, AZCORR*1000, ELCORR*1000, TOF, VP
240 PRINT #2, T, AZCORR*1000
250 PRINT #3, T, ELCORR*1000
260 PRINT #4, T, AZCORR*1000, ELCORR*1000
280 GOTO 100
```

COMPUTER PROGRAM: GUNDATPR.BAS

PURPOSE: Computes Azimuth and Elevation Angles along the Gun Boresight Axis. Uses Equation IX-4 and IX-5.

RESULTS: Used in GUNTHEPR.BAS and GUNCORPR.BAS

```
 10 OPEN "I", #1, "B:LOSDATA.CS5"
20 OPEN "O", #2, "B:gundata.CS5"
30 PI = 3.1417
40 G = 32.17
50 VMUZ = 2840
60 INPUT "Aircraft Velocity = ?", VAC
70 VAC = VAC*88/60/.8684
80 INPUT #1, T, AZLAS, ELLAS, RSL
90 VP = SQR(VMUZ^2 + 2*VMUZ*VAC*SIN(AZLAS)*COS(ELLAS) + VAC^2)
100 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
110 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
120 RHO = .07479
130 C = 1.5*RHO/WCDA
140 TOF = RSL/VP/(1 - C*RSL/4)^2
150 ELGD = 16.085*COS(ELLAS)*(BETA*TOF)^2/RSL
160 ELCORR = ELGD
170 AZGUN = VP*SIN(AZLAS)*COS(ELLAS) - VAC
180 AZGUN = AZGUN/VMUZ/COS(ELGD + ELLAS)
190 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
200 ELGUN = ELLAS + ELGD
210 PRINT T, AZGUN*180/PI, ELGUN*180/PI, RSL
220 PRINT #2, T, AZGUN, ELGUN, RSL
230 GOTO 80
```

COMPUTER PROGRAM: GUNTHEPR.BAS

PURPOSE: Computes the Azimuth and Elevation Corrections Based on Azimuth and Elevation Angle Measurements Made along Gun Boresight Axis

RESULTS: Figures IX-3 through IX-12

```
010 OPEN "I", #1, "B:GUNDATA.CS5"
20 OPEN "O", #2, "B:GUNAZTHE.CS5"
30 OPEN "O", #3, "B:GUNELTHE.CS5"
40 OPEN "O", #4, "B:OUTTHE.CS5"
50 PI = 3.1417
60 G = 32.17
70 VMUZ = 2840
80 INPUT "Aircraft Velocity = ?", VAC
90 VAC = VAC*88/60/.8684
100 INPUT #1, T, AZLAS, ELLAS, RSL
110 VP = SQR(VMUZ^2 + 2*VMUZ*VAC*SIN(AZLAS)*COS(ELLAS) + VAC^2)
120 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
130 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
140 RHO = .07479
150 C = 1.5*RHO/WCDA
160 TOF = RSL/VP/(1 - C*RSL/4)^2
170 ELGD = 16.085*COS(ELLAS)*(BETA*TOF)^2/RSL
180 ELCORR = ELGD
190 AZGUN = VP*SIN(AZLAS)*COS(ELLAS) - VAC
200 AZGUN = AZGUN/VMUZ/COS(ELGD + ELLAS)
210 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
220 AZCORR = AZLAS - AZGUN
230 PRINT T, AZLAS*180/PI, AZCORR*1000, ELCORR*1000, TOF, VP
240 PRINT #2, T, AZCORR*1000
250 PRINT #3, T, ELCORR*1000
260 PRINT #4, T, AZCORR*1000, ELCORR*1000
270 GOTO 100
```

COMPUTER PROGRAM: GUNCORPR.BAS

PURPOSE: Computes the Azimuth and Elevation Corrections Based on Azimuth and Elevation Angle Measurements Made along Gun Boresight Axis, but Corrects these Measurements using the Previous Correction before Making the Computation.

RESULTS: Figures IX-3 through IX-12

```
010 OPEN "I", #1, "B:GUNDATA.CS5"
20 OPEN "O", #2, "B:GUNAZCOR.CS5"
30 OPEN "O", #3, "B:GUNELCOR.CS5"
40 OPEN "O", #4, "B:OUTCOR.CS5"
50 PI = 3.1417
60 G = 32.17
70 VMUZ = 2840
80 INPUT "Aircraft Velocity = ?", VAC
90 VAC = VAC*88/60/.8684
100 INPUT #1, T, AZLAS, ELLAS, RSL
110 AZLAS = AZLAS + AZCORR
120 ELLAS = ELLAS - ELCORR
130 VP = SQR(VMUZ^2 + 2*VMUZ*VAC*SIN(AZLAS)*COS(ELLAS) + VAC^2)
140 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
150 BETA = 3E-09*RSL^2 - 4.706E-05*RSL + 1.0216
160 RHO = .07479
170 C = 1.5*RHO/WCDA
180 TOF = RSL/VP/(1 - C*RSL/4)^2
190 ELGD = 16.085*COS(ELLAS)*(BETA*TOF)^2/RSL
200 ELCORR = ELGD
210 AZGUN = VP*SIN(AZLAS)*COS(ELLAS) - VAC
220 AZGUN = AZGUN/VMUZ/COS(ELGD + ELLAS)
230 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
240 AZCORR = AZLAS - AZGUN
250 PRINT T, AZLAS*180/PI, AZCORR*1000, ELCORR*1000, TOF, VP
260 PRINT #2, T, AZCORR*1000
270 PRINT #3, T, ELCORR*1000
280 PRINT #4, T, AZCORR*1000, ELCORR*1000
290 GOTO 100
```

0

COMPUTER PROGRAM: CORRECTP.BAS

PURPOSE: (a) Computes Nominal Slant Range and Azimuth/Elevation Angles for the LOS; (b) Adds Noise to Nominal Slant Range and Azimuth/Elevation Angles; (c) Computes Noisy Gun Boresight Azimuth/Elevation Angles; (d) Filters Range and Azimuth/Elevation Measurements; (e) Computes Azimuth/Elevation Aiming Corrections; (f) Computes Error in the Aiming Correction.

RESULTS: Error in Azimuth and Elevation Aiming Corrections, as well as Individual Contributions to these Errors.

```
 10 OPEN "I", #1, "C:NOISEDAT.DAT"
20 OPEN "O", #2, "B:AZCORRQ.C01"
30 OPEN "O", #3, "B:AZTOTDIF.C01"
40 OPEN "O", #4, "B:ELTOTDIF.C01"
50 OPEN "O", #5, "B:ELCORRDI.c01"
60 RHO = .07479
65 G = 32.17
70 DT = .017
80 VMUZ = 2840
90 PI = 3.1416
100 KPR = .5
110 KVR = .5
120 KPA = .05
130 KVA = .05
140 KPE = .05
150 KVE = .05
160 INPUT "Aircraft Altitude (feet) = ?", H
170 INPUT "Aircraft Nominal Velocity (knots) = ?", VEL
180 VAC = VEL*88/60/.8684
190 INPUT "Range at Closest Approach (meters) =?", RMIN
200 RMIN = RMIN*3.28
210 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
220 GOSUB 940
240 GOSUB 1150
250 PRINT #2, T, AZCORR*1000
260 PRINT #4, T, ELGD*1000
270 RHAT = RSLN
280 RHAT1 = RHAT
290 VACHAT = VAC
300 AZHAT = AZLOSN
310 AZHAT1 = AZHAT
320 ELHAT = ELLOSN
```

```
330 ELHAT1 = ELLOSN
350 GOSUB 1250
360 AZDIFF = AZCORR - AZCORRN
370 ELDIFF = ELGD - ELGDN
371 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
380 PRINT #3, T, AZDIFF*1000
390 PRINT #5, T, ELDIFF*1000
400 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
420 GOSUB 940
430 GOSUB 1150
440 PRINT #2, T, AZCORR*1000
450 PRINT #4, T, ELGD*1000
460 RHAT = RHAT + (RSLN - RHAT)/2
470 RDHAT =  (RHAT - RHAT1)/DT
480 RHAT1 = RHAT
490 VACHAT = VACHAT + (VACN - VACHAT)/2
500 AZHAT = AZHAT + (AZLOSN - AZHAT)/2
510 AZDHAT = (AZHAT - AZHAT1)/DT
520 AZHAT1 = AZHAT
530 ELHAT = ELHAT + (ELLOSN - ELHAT)/2
540 ELDHAT = (ELHAT - ELHAT1)/DT
550 ELHAT1 = ELHAT
560 GOSUB 1250
565 AZDIFF = AZCORR - AZCORRN
566 ELDIFF = ELGD - ELGDN
570 PRINT #3, T, AZDIFF*1000
580 PRINT #5, T, ELDIFF*1000
581 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
600 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
610 GOSUB 940
620 GOSUB 1150
630 PRINT #2, T, AZCORR*1000
640 PRINT #4, T, ELGD*1000
650 RHAT = RHAT + (RSLN - RHAT)/3
660 RDHAT =  (RHAT - RHAT1)/DT
670 RHAT1 = RHAT
680 VACHAT = VACHAT + (VACN - VACHAT)/3
690 AZHAT = AZHAT + (AZLOSN - AZHAT)/3
700 AZDHAT = (AZHAT - AZHAT1)/DT
710 AZHAT1 = AZHAT
720 ELHAT = ELHAT + (ELLOSN - ELHAT)/3
730 ELDHAT = (ELHAT - ELHAT1)/DT
740 ELHAT1 = ELHAT
750 GOSUB 1250
755 AZDIFF = AZCORR - AZCORRN
756 ELDIFF = ELGD - ELGDN
760 PRINT #3, T, AZDIFF*1000
770 PRINT #5, T, ELDIFF*1000
771 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
790 K = 3
800 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
810 GOSUB 940
```

```
820 GOSUB 1150
830 PRINT #2, T, AZCORR*1000
840 PRINT #4, T, ELGD*1000
850 GOSUB 1070
860 VACHAT = VACHAT + (VACN - VACHAT)/(K +1)
870 GOSUB 1250
875 AZDIFF = AZCORR - AZCORRN
876 ELDIFF = ELGD - ELGDN
880 PRINT #3, T, AZDIFF*1000
890 PRINT #5, T, ELDIFF*1000
891 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
900 IF AZLOS < 0!  THEN GOTO 1350
910 T = T + .017
920 K = K + 1
930 GOTO 800
940 RSL = SQR(H^2 + RMIN^2 + (RMIN - VAC*T)^2)
950 RSLN = RSL + RGNOISE
960 ELLOS = H/RSL
970 ELLOS = ATN(ELLOS/SQR(1 - ELLOS^2))
980 ELLOSN = ELLOS + ELNOISE/RSL
990 AZLOS = ATN((RMIN - VAC*T)/RMIN)
1000 AZLOSN = AZLOS + AZNOISE/RSL
1010 VP = VMUZ^2 + 2*VMUZ*VAC*SIN(AZLOS)*COS(ELLOS) + VAC^2
1020 VP = SQR(VP)
1030 VACN = VAC - RGNOISE*VAC/460
1040 VPN = VMUZ^2 + 2*VMUZ*VACN*SIN(AZLOSN)*COS(ELLOSN) + VACN^2
1050 VPN =SQR(VPN)
1060 RETURN
1070 OMEGA = SQR(AZDHAT^2 + ELDHAT^2)
1080 RHAT = (1 - KPR)*RHAT + DT*(1 - KPR)*RDHAT + KPR*RSLN
1090 RDHAT = (DT*OMEGA^2 - KVR)*RHAT + (1 - DT*KVR)*RDHAT + KVR*RSLN
1100 AZHAT = (1 - KPA)*AZHAT + KPA*AZLOSN + DT*(1 - KPA)*AZDHAT
1110 AZDHAT = -KVA*AZHAT + (1 - DT*KVA - 2*RDHAT*DT/RHAT)*AZDHAT + KVA*AZLOSN
1120 ELHAT = (1 - KPE)*ELHAT + KPE*ELLOSN + DT*(1 - KPE)*ELDHAT
1130 ELDHAT = -KVE*ELHAT + (1 - DT*KVE - 2*RDHAT*DT/RHAT)*ELDHAT + KVE*ELLOSN
1140 RETURN
1150 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
1160 BETA = 3E-09*RSL^2 - 4.0706E-05*RSL + 1.0216
1170 C = 1.5*RHO/WCDA
1180 TOF = RSL/VP/(1 - C*RSL/4)^2
1190 ELGD = .5*G*COS(ELLOS)*(BETA*TOF)^2/RSL
1200 AZGUN = VP*SIN(AZLOS)*COS(ELLOS) - VAC
1210 AZGUN = AZGUN/VMUZ/COS(ELLOS + ELGD)
1220 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
1230 AZCORR = AZLOS - AZGUN
1240 RETURN
1250 WCDA = .000005*RHAT^2 - .036938*RHAT + 700.4299
1260 BETA = 3E-09*RHAT^2 - 4.0706E-05*RHAT + 1.0216
1270 C = 1.5*RHO/WCDA
1280 TOF = RHAT/VPN/(1 - C*RHAT/4)^2
1290 ELGDN = .5*G*COS(ELHAT)*(BETA*TOF)^2/RHAT
1300 AZGUNN = VPN*SIN(AZHAT)*COS(ELHAT) - VACN
1310 AZGUNN = AZGUNN/VMUZ/COS(ELHAT + ELGDN)
1320 AZGUNN = ATN(AZGUNN/SQR(1 - AZGUNN^2))
1330 AZCORRN = AZHAT - AZGUNN
1340 RETURN
1350 END
```

CHAPTER X

FIRE CONTROL SYSTEM USING EXTERNAL MEASUREMENTS

Analysis with Simulated Noisy Data

A fire control system that uses external measurement of aircraft velocity, a laser rangefinder, gun boresight aligned angular resolvers for azimuth/elevation LOS and an inertial platform was detailed in the previous chapter. The previous analysis showed that by adding back the previous azimuth and elevation correction to the measurements made along the gun boresight, the system performs essentially as if measurements were made along the laser beam axis as required.

However, the previous analysis of this system did not use "noisy" measurements that result when a gunner is aiming the laser beam via the gun mounted in a moving helicopter. Thus in this analysis "noisy" range and angular measurements were included to provide a more realistic operating environment.

The noisy range data was based on laser range measurements actually measured during the BSTING flight test

program conducted during August and September of 1990. The noisy range data was the same as used in the previous investigations to evaluate improvements to the BSTING system. Figure X-1 shows the zero-mean noise added to the nominal range data.

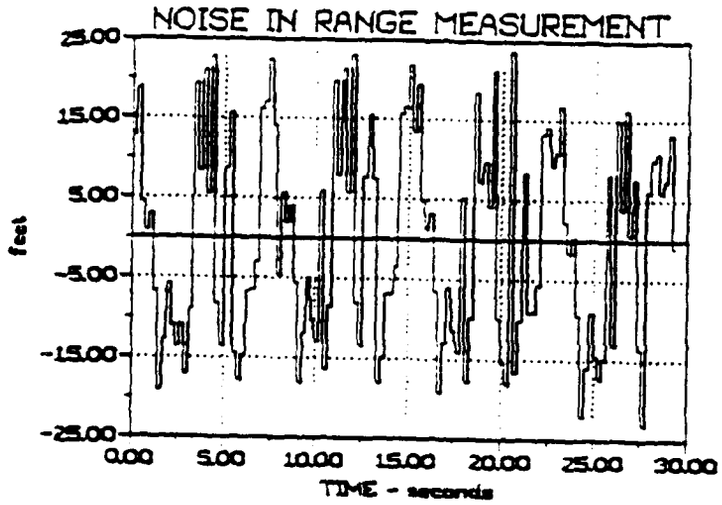


Figure X-1

Zero-Mean Range Noise

During this same test program, the location of the laser spot on the target was measured by a low light level TV camera. This data was used here to determine a representative noise profile for measurements made by the azimuth and elevation angular resolvers mounted on the gun. The noise extracted from the flight test data for a typical pass is shown in figures X-2 and X-3. It was assumed that the gunner kept the spot within the approximately 40 ft x 40 ft area of the target and thus the noise data was "faired" in during the period when the laser beam was off and thus its position was not recorded. Like the range noise, azimuth/elevation noise was normalized to have a zero mean value. These azimuth/elevation noise measurements were then divided by the instantaneous nominal slant range to obtain angular noise values. These values were then added to the nominal azimuth and elevation angles determined in the previous chapter.

In this method, the aircraft velocity is supplied from an external source. It was assumed that velocity contains a certain amount of noise. This noise was modeled by assumimg the noise was 5-percent of the noise in range measurements and of opposite sign, i.e.

$$VELN = (1 - 0.05\ NOISE_{range})$$

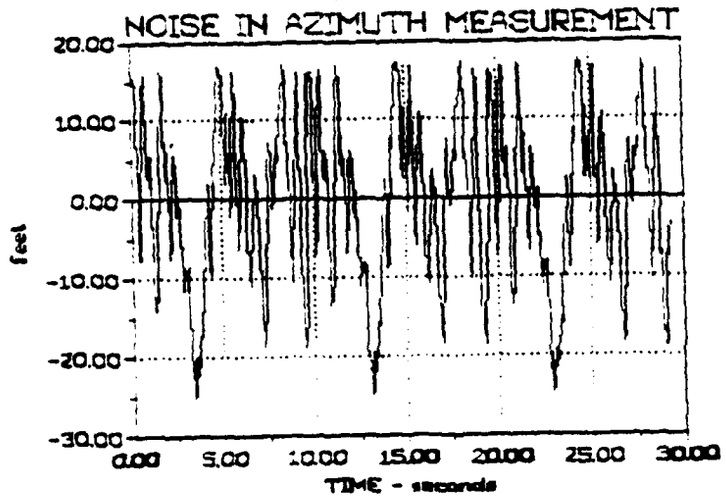


Figure X-2

Zero Mean Horizontal Laser Spot Location

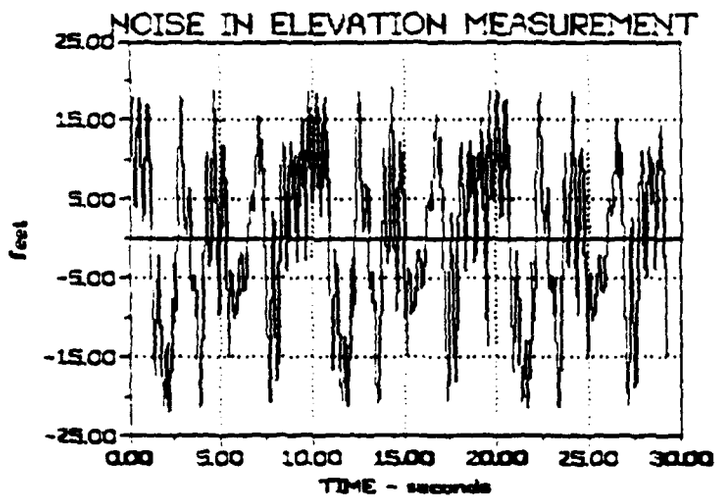


Figure X-3

Zero Mean Vertical Laser Spot Location

Since azimuth and elevation angles rather than azimuth and elevation rates are being measured, a new set of filters were needed to smooth the noisy range and azimuth/elevation data. These are described here;

Range Filter

```
RHAT(k+1) = (1 - KPR)*RHAT(k) + (1 - KPR)*RDHAT(k+1)*DT
          + KPR*RSLN

RDHAT(k+1) = (OMEGA^2*DT - KVR)*RHAT(k) + (1 - KPR*DT)*RDHAT(k)
           + KVR*RSLN
```

Azimuth Angle Filter

```
AZHAT(k+1) = (1 - KPR)*AZHAT(k) + (1 - KPR)*AZDHAT(k+1)*DT
           + KPR*AZGUN

AZDHAT(k+1) = -KVA*AZHAT(k) + (1 - KVA)*DT
            - [2*DT*RDHAT(k)/RHAT(k)]*AZDHAT(k) + KVA*AZLOSN
```

Elevation Angle Filter

```
ELHAT(k+1) = (1 - KPE)*ELHAT(k) + (1 - KPE)*ELDHAT(k+1)*DT
           + KPE*ELGUN

ELDHAT(k+1) = -KVE*ELHAT(k) + (1 - KVE)*DT
            -[2*DT*RDHAT(k)/RHAT(k)]*ELDHAT(k) + KVE*ELLOSN
```

WHERE: $OMEGA^2 = AZDHAT(k)^2 + ELDHAT(k)^2$

RSLN = Noisy Range Data

AZLOSN = Noisy Azimuth Data (Measured along Gun Boresight then Corrected Using Previous Azimuth Correction Value.)

ELLOSN = Noisy Elevation Data (Measured along Gun Boresight then Corrected Using Previous Elevation Correction Value.)

DT = Time increment between noisy data samples.

<u>Aircraft Velocity Filter</u>

$$VACHAT(k+1) = VACHAT(k) + [VACN - VACHAT(k)/(k + 1)$$

The filtered data was then used in the following algorithms to determine the azimuth and elevation corrections.

<u>Elevation Angle Correction</u>

$$ELGD = 0.5*COS(ELHAT)*(BETA*TOF)/RHAT$$

$$TOF = \frac{RHAT}{VPN*(1 - c*RHAT/4)^2}$$

$$c = 1.5*density/WCDA$$

$$WCDA = .00005*RHAT^2 - 0.036938*RHAT + 700.4299$$

$$Beta = 3E-09*RHAT^2 - 4.706E-05*RHAT + 1.0216$$

$$density = 0.07479$$

$$ELCORR = ELGD$$

Azimuth Angle Correction

$$\text{Sin (AZGUN)} = \frac{\text{VP*Sin(AZHAT)*COS(ELHAT)} - \text{Va/c}}{\text{Vmuz *Cos(ELLOSN + ELGD)}}$$

$$VPN^2 = V_{muz}^2 + 2*V_{muz}*V_{a/c}*SIN(AZLOSN)*COS(ELLOSN) + V_{a/c}^2$$

$$AZCORR = AZLOSN - AZGUN$$

Starting Values

For each parameter, the first three values were computed by:

$$"X"HAT(k+1) = "X"HAT(k) + ["X"N - "X"HAT(k]/(k + 1)$$

Where: k= 0, 1, 2

Starting values for rate parameters were determined by:

$$"X"DHAT(k+1) = ["X"HAT(k+1) - "X"HAT(k)]/dt$$

Only one scenario was used in this investigation and was:

Va/c = 150 knots (helicopter velocity)

Rmin = 1000 meter (range at closest approach)

H = 250 feet (altitude above target)

A brief investigation was done to determine the "best" values for the filtering coefficients - KPR, KVR, KPA, KPE, KVA and KPE. The results are shown in figures X-4 through X-9 in terms of the difference between the nominal and filtered noisy data. From this analysis, the following coefficients values were used in subsequent analysis.

KVR	=	0.5	KPR	=	0.5
KPA	=	0.05	KVA	=	0.05
KVE	=	0.05			

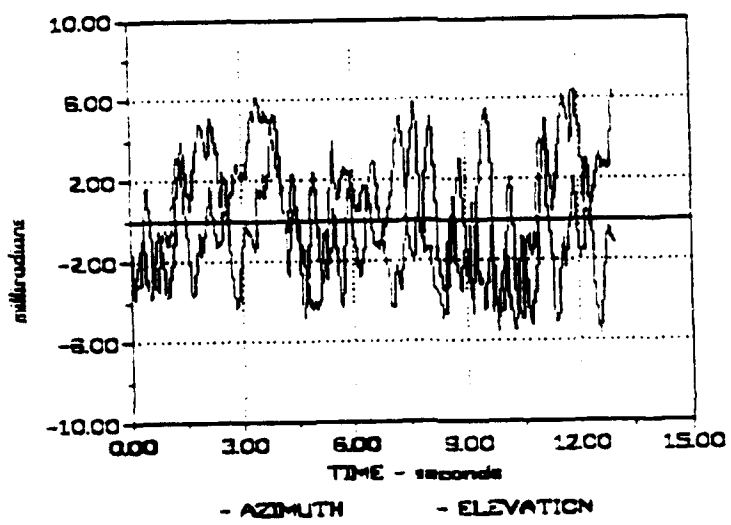


Figure X-4

Unfiltered Differences between Nominal and Noisy Azimuth and Elevation Angle Measurements

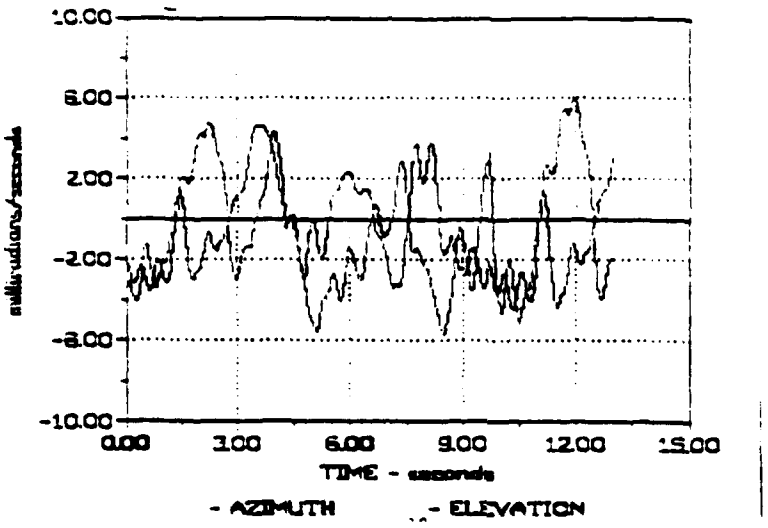


Figure X-5

Filtered Differences between Nominal and Noisy Azimuth and Elevation Angle Measurements. (KPA = KPR = KPE = KVE = 0.1)

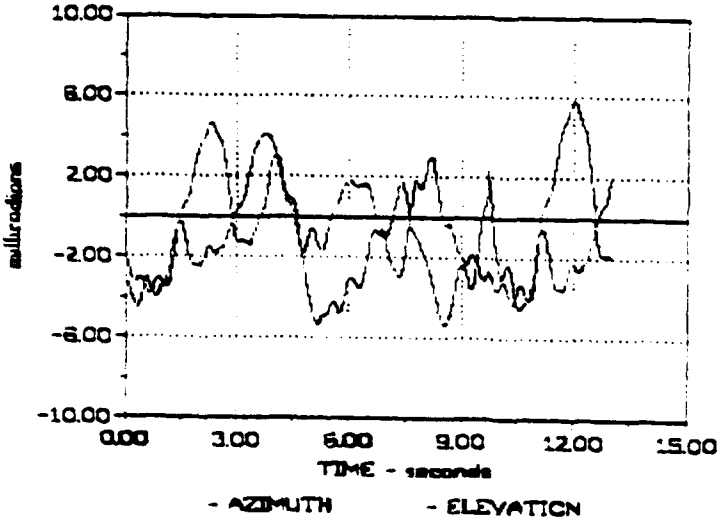


Figure X-6

Filtered Differences between Nominal and Noisy Azimuth and Elevation Angle Measurements. (KPA = KPR = KPE = KVE = 0.05)

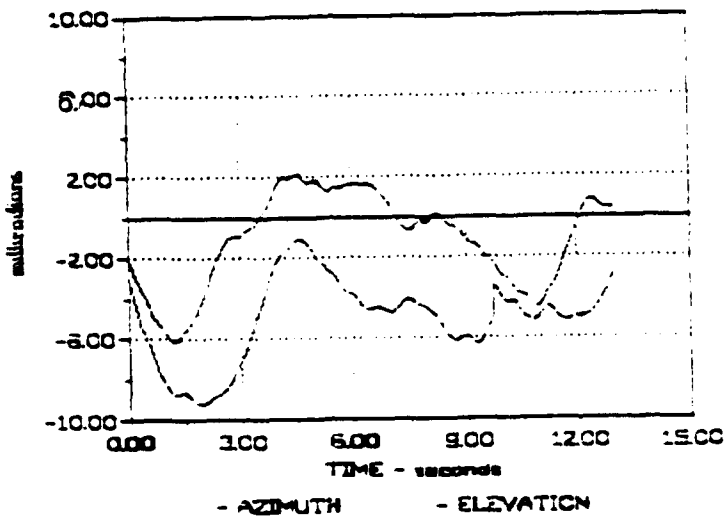


Figure X-7

Filtered Differences between Nominal and Noisy Azimuth and Elevation Angle Measurements. (KPA = KPR = KPE = KVE = 0.01)

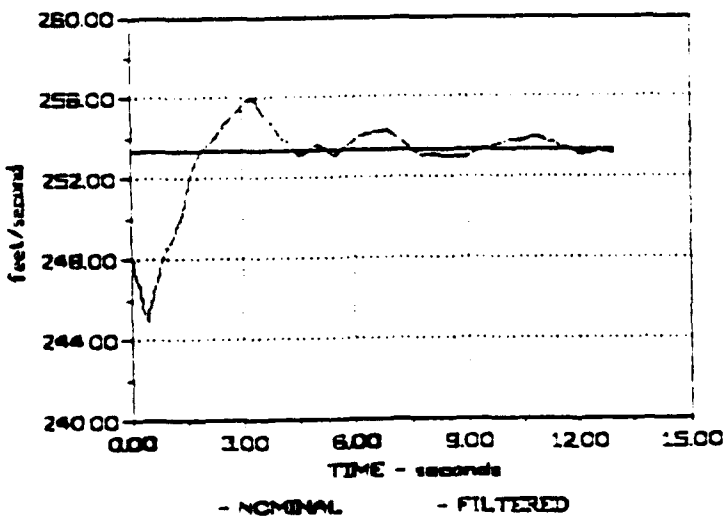


Figure X-8

Filtered Aircraft Velocity

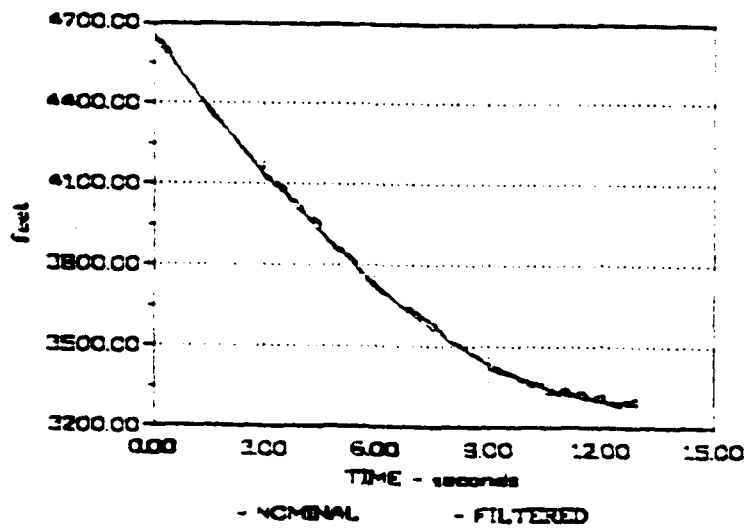


Figure X-9

Filtered Slant Range (KVR = KPR = 0.5)

The results of this investigation are shown in figure X-10 and X-11 in the form of the computed azimuth and elevation corrections based on the noisy data versus the required corrections for accurate aiming. Because of the large sampling rate, the azimuth range used started at 45 degrees off target and ended at 0 degrees with respect to the target.

For this fire control system, there are four measurements that contribute to the total error. These are the errors in slant range, aircraft velocity, LOS azimuth and LOS elevation angles. This system also requires an inertial platform, but the error contribution of the inertial platform were not included in this initial investigation.

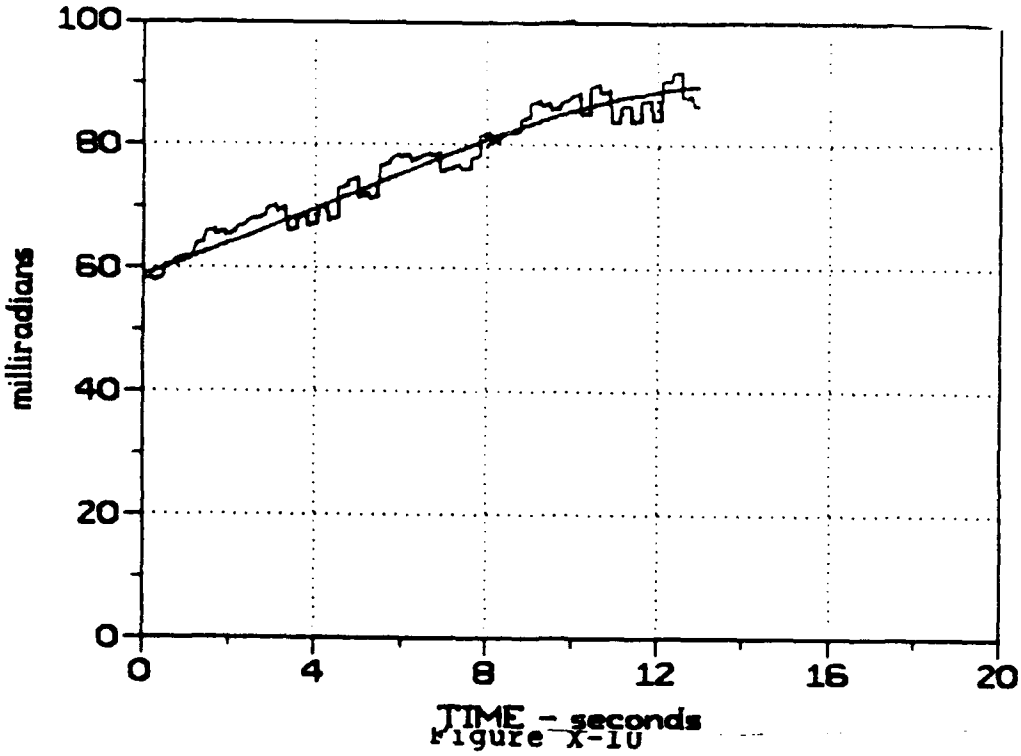


Figure X-10

Total Correction in Azimuth Plane - Computed versus Required

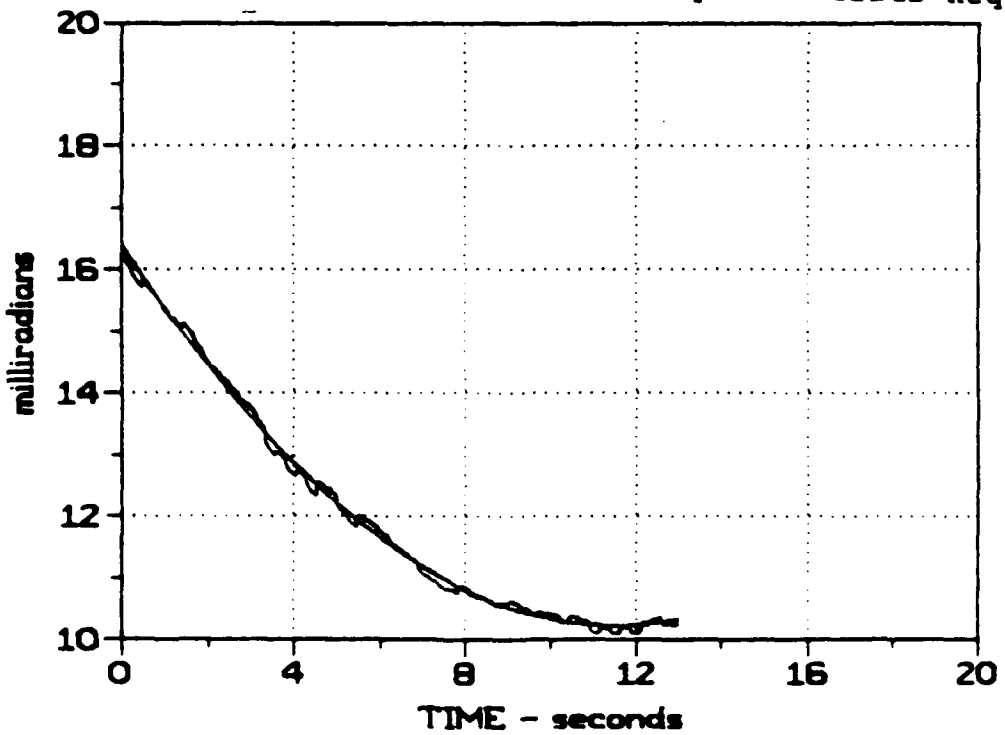


Figure X-11

Total Correction in Elevation Plane - Computed versus Required

These individual contributions are shown as well as the total error are shown in figures X-12 and X-13 for the azimuth and elevation aiming corrections respectively. In both cases, the graphs show the difference between the computed corrections and the required corrections. By far, the largest error comes from errors in the aircraft velocity. This can also be discerned from the algorithm used to determine the azimuth correction. Incidently, in this investigation no error in vertical velocity was assumed. In a real situation, there would be some error in vertical velocity and consequently an increase in elevation aiming error.

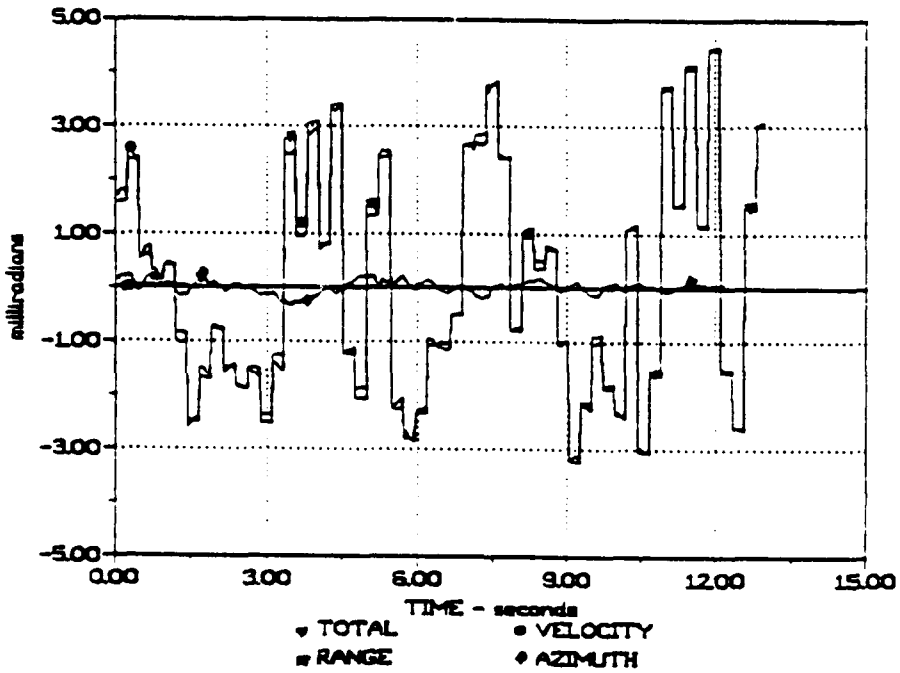


Figure X-12

Error Contributions - Azimuth Plane

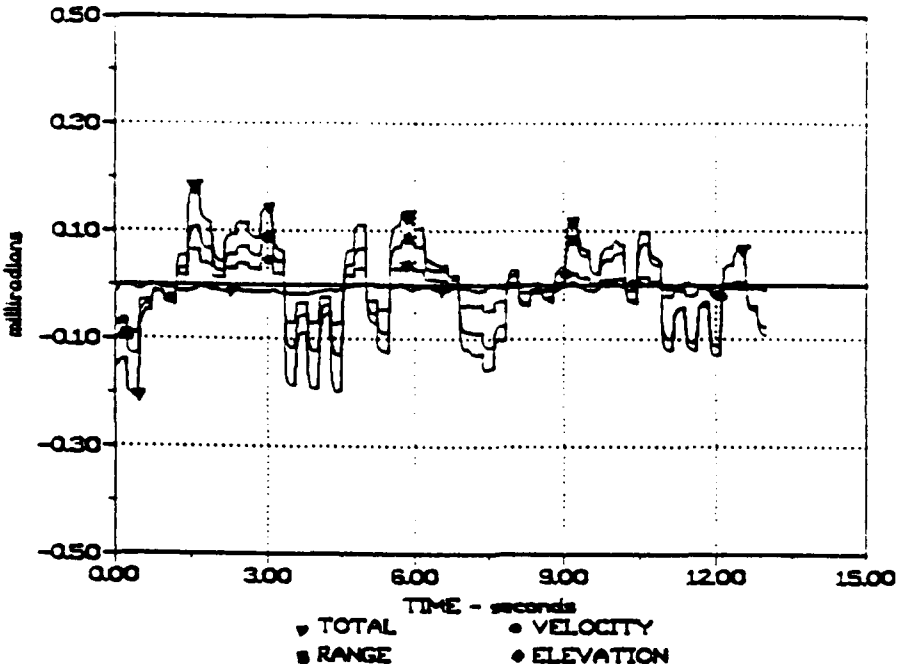


Figure X-13

Error Contributions - Elevation Plane

Since aircraft velocity is the largest contributor to the total azimuth aiming error, an investigation was made into the sensitivity of total aiming errors in aircraft velocity measurement. Again the same scenario was used and two azimuth pointing directions were used - 45, and 0 degrees off of the RMIN (range of closest approach). The results are shown in figure X-14 and X-15.

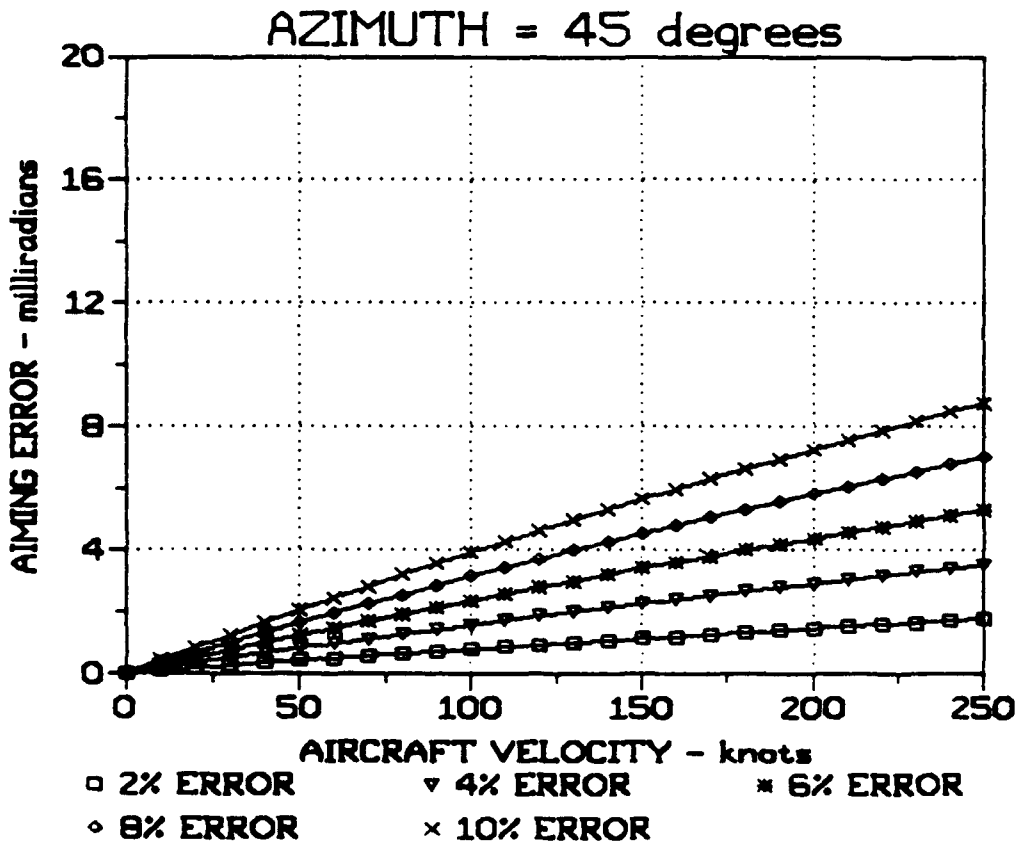


Figure X-14

Sensitivity of the Total Azimuth Aiming Error to Errors in Aircraft Velocity Measurements

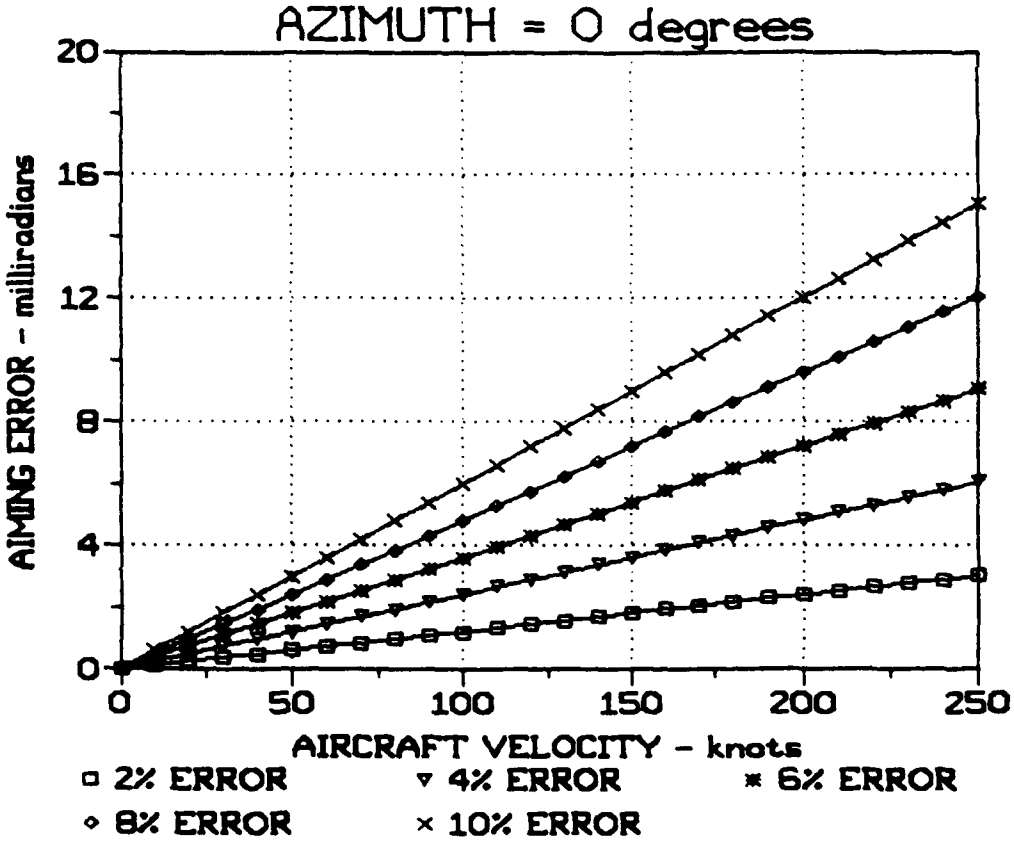


Figure X-15

Sensitivity of the Total Elevation Aiming Error to Errors in Aircraft Velocity Measurements

CONCLUSIONS:

While this fire control system does reduce the sensitivity to azimuth and elevation errors due to gunner pointing difficulties, the system requires requires very accurate measurement of the aircraft velocity. From figure X-15, it can be concluded that aircraft velocity would have to measured at better than 2 percent.

Also not investigated in this study, but required in an actual system are the complexity and inaccuracies of inertial platforms and altitude measurements that also would have to be used with this system.

COMPUTER PROGRAM: CORRECTP.BAS

PURPOSE: (a) Computes Nominal Slant Range and Azimuth/Elevation Angles for the LOS; (b) Adds Noise to Nominal Slant Range and Azimuth/Elevation Angles; (c) Computes Noisy Gun Boresight Azimuth/Elevation Angles; (d) Filters Range and Azimuth/Elevation Measurements; (e) Computes Azimuth/Elevation Aiming Corrections; (f) Computes Error in the Aiming Correction.

RESULTS: Error in Azimuth and Elevation Aiming Corrections, as well as Individual Contributions to these Errors.

```
 10 OPEN "I", #1, "C:NOISEDAT.DAT"
20 OPEN "O", #2, "B:AZCORR0.C01"
30 OPEN "O", #3, "B:AZTOTDIF.C01"
40 OPEN "O", #4, "B:ELTOTDIF.C01"
50 OPEN "O", #5, "B:ELCORRDI.c01"
60 RHO = .07479
65 G = 32.17
70 DT = .017
80 VMUZ = 2840
90 PI = 3.1416
100 KPR = .5
110 KVR = .5
120 KPA = .05
130 KVA = .05
140 KPE = .05
150 KVE = .05
160 INPUT "Aircraft Altitude (feet) = ?", H
170 INPUT "Aircraft Nominal Velocity (knots) = ?", VEL
180 VAC = VEL*88/60/.8684
190 INPUT "Range at Closest Approach (meters) =?", RMIN
200 RMIN = RMIN*3.28
210 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
220 GOSUB 940
240 GOSUB 1150
250 PRINT #2, T, AZCORR*1000
260 PRINT #4, T, ELGD*1000
270 RHAT = RSLN
280 RHAT1 = RHAT
290 VACHAT = VAC
300 AZHAT = AZLOSN
310 AZHAT1 = AZHAT
320 ELHAT = ELLOSN
```

COMPUTER PROGRAM: VELSENSI.BAS

PURPOSE: Computes the Azimuth Aiming Error that Results from Errors in Measuring Aircraft Velocity. Allows Different Percentages of Error and Aircraft Velocities from 0 to 250 knots.

RESULTS: Azimuth Aiming Errors as a Function of Aircraft Velocity and Percentage Error in Velocity.

```
010 OPEN "O", #1, "B:VEL10%.C45"
20 PERCENT = .1
30 PI = 3.1417
40 G = 32.17
50 VMUZ = 2840
60 INPUT "LOS Azimuth wrt Platform (deg) = ?", AZ
70 INPUT "LOS Elevation wrt Platform (deg) = ?", EL
80 AZ1 = AZ*PI/180
90 EL1 = EL*PI/180
100 VAC = VEL*88/60/.8684
110 GOSUB 210
120 AZCORRTH = AZCORR
130 VAC = VAC*(1 + PERCENT)
140 GOSUB 210
150 AZCORRERR = AZCORR
160 AZDIFF = -(AZCORRTH - AZCORRERR)
170 PRINT VEL, AZCORRTH*1000, AZCORRERR*1000, AZDIFF*1000
175 PRINT #1, VEL, AZDIFF*1000
180 IF VEL > 240 THEN GOTO 280
190 VEL = VEL + 10
200 GOTO 100
210 VP = VMUZ^2 + 2*VMUZ*VAC*SIN(AZ1)*COS(EL1) + VAC^2
220 VP = SQR(VP)
230 AZGUN = VP*SIN(AZ1)*COS(EL1) - VAC
240 AZGUN = AZGUN/VMUZ/COS(EL1)
250 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
260 AZCORR = AZ1 - AZGUN
270 RETURN
280 END
```

```
330 ELHAT1 = ELLOSN
350 GOSUB 1250
360 AZDIFF = AZCORR - AZCORRN
370 ELDIFF = ELGD - ELGDN
371 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
380 PRINT #3, T, AZDIFF*1000
390 PRINT #5, T, ELDIFF*1000
400 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
420 GOSUB 940
430 GOSUB 1150
440 PRINT #2, T, AZCORR*1000
450 PRINT #4, T, ELGD*1000
460 RHAT = RHAT + (RSLN - RHAT)/2
470 RDHAT =  (RHAT - RHAT1)/DT
480 RHAT1 = RHAT
490 VACHAT = VACHAT + (VACN - VACHAT)/2
500 AZHAT = AZHAT + (AZLOSN - AZHAT)/2
510 AZDHAT = (AZHAT - AZHAT1)/DT
520 AZHAT1 = AZHAT
530 ELHAT = ELHAT + (ELLOSN - ELHAT)/2
540 ELDHAT = (ELHAT - ELHAT1)/DT
550 ELHAT1 = ELHAT
560 GOSUB 1250
565 AZDIFF = AZCORR - AZCORRN
566 ELDIFF = ELGD - ELGDN
570 PRINT #3, T, AZDIFF*1000
580 PRINT #5, T, ELDIFF*1000
581 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
600 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
610 GOSUB 940
620 GOSUB 1150
630 PRINT #2, T, AZCORR*1000
640 PRINT #4, T, ELGD*1000
650 RHAT = RHAT + (RSLN - RHAT)/3
660 RDHAT =  (RHAT - RHAT1)/DT
670 RHAT1 = RHAT
680 VACHAT = VACHAT + (VACN - VACHAT)/3
690 AZHAT = AZHAT + (AZLOSN - AZHAT)/3
700 AZDHAT = (AZHAT - AZHAT1)/DT
710 AZHAT1 = AZHAT
720 ELHAT = ELHAT + (ELLOSN - ELHAT)/3
730 ELDHAT = (ELHAT - ELHAT1)/DT
740 ELHAT1 = ELHAT
750 GOSUB 1250
755 AZDIFF = AZCORR - AZCORRN
756 ELDIFF = ELGD - ELGDN
760 PRINT #3, T, AZDIFF*1000
770 PRINT #5, T, ELDIFF*1000
771 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
790 K = 3
800 INPUT #1, T, AZNOISE, ELNOISE, RGNOISE
810 GOSUB 940
```

```
820 GOSUB 1150
830 PRINT #2, T, AZCORR*1000
840 PRINT #4, T, ELGD*1000
850 GOSUB 1070
860 VACHAT = VACHAT + (VACN - VACHAT)/(K +1)
870 GOSUB 1250
875 AZDIFF = AZCORR - AZCORRN
876 ELDIFF = ELGD - ELGDN
880 PRINT #3, T, AZDIFF*1000
890 PRINT #5, T, ELDIFF*1000
891 PRINT T, AZLOS*180/PI, AZDIFF*1000, VAC
900 IF AZLOS < 0!  THEN GOTO 1350
910 T = T + .017
920 K = K + 1
930 GOTO 800
940 RSL = SQR(H^2 + RMIN^2 + (RMIN - VAC*T)^2)
950 RSLN = RSL + RGNOISE
960 ELLOS = H/RSL
970 ELLOS = ATN(ELLOS/SQR(1 - ELLOS^2))
980 ELLOSN = ELLOS + ELNOISE/RSL
990 AZLOS = ATN((RMIN - VAC*T)/RMIN)
1000 AZLOSN = AZLOS + AZNOISE/RSL
1010 VP = VMUZ^2 + 2*VMUZ*VAC*SIN(AZLOS)*COS(ELLOS) + VAC^2
1020 VP = SQR(VP)
1030 VACN = VAC - RGNOISE*VAC/460
1040 VPN = VMUZ^2 + 2*VMUZ*VACN*SIN(AZLOSN)*COS(ELLOSN) + VACN^2
1050 VPN =SQR(VPN)
1060 RETURN
1070 OMEGA = SQR(AZDHAT^2 + ELDHAT^2)
1080 RHAT = (1 - KPR)*RHAT + DT*(1 - KPR)*RDHAT + KPR*RSLN
1090 RDHAT = (DT*OMEGA^2 - KVR)*RHAT + (1 - DT*KVR)*RDHAT + KVR*RSLN
1100 AZHAT = (1 - KPA)*AZHAT + KPA*AZLOSN + DT*(1 - KPA)*AZDHAT
1110 AZDHAT = -KVA*AZHAT + (1 - DT*KVA - 2*RDHAT*DT/RHAT)*AZDHAT + KVA*AZLOSN
1120 ELHAT = (1 - KPE)*ELHAT + KPE*ELLOSN + DT*(1 - KPE)*ELDHAT
1130 ELDHAT = -KVE*ELHAT + (1 - DT*KVE - 2*RDHAT*DT/RHAT)*ELDHAT + KVE*ELLOSN
1140 RETURN
1150 WCDA = .000005*RSL^2 - .036938*RSL + 700.4299
1160 BETA = 3E-09*RSL^2 - 4.0706E-05*RSL + 1.0216
1170 C = 1.5*RHO/WCDA
1180 TOF = RSL/VP/(1 - C*RSL/4)^2
1190 ELGD = .5*G*COS(ELLOS)*(BETA*TOF)^2/RSL
1200 AZGUN = VP*SIN(AZLOS)*COS(ELLOS) - VAC
1210 AZGUN = AZGUN/VMUZ/COS(ELLOS + ELGD)
1220 AZGUN = ATN(AZGUN/SQR(1 - AZGUN^2))
1230 AZCORR = AZLOS - AZGUN
1240 RETURN
1250 WCDA = .000005*RHAT^2 - .036938*RHAT + 700.4299
1260 BETA = 3E-09*RHAT^2 - 4.0706E-05*RHAT + 1.0216
1270 C = 1.5*RHO/WCDA
1280 TOF = RHAT/VPN/(1 - C*RHAT/4)^2
1290 ELGDN = .5*G*COS(ELHAT)*(BETA*TOF)^2/RHAT
1300 AZGUNN = VPN*SIN(AZHAT)*COS(ELHAT) - VACN
1310 AZGUNN = AZGUNN/VMUZ/COS(ELHAT + ELGDN)
1320 AZGUNN = ATN(AZGUNN/SQR(1 - AZGUNN^2))
1330 AZCORRN = AZHAT - AZGUNN
1340 RETURN
1350 END
```

CHAPTER XI

FURTHER IMPROVEMENTS FOR CONTINUOUS TRACKING AND AIMING

Adaptive Filters and Increased Sampling Rates

Three promising techniques for improving the BSTING-type fire control system so that it provides more accurate, continuous tracking and aiming as well as rapid response approaching a "point and shoot" capability have been discussed previously. These are (a) the gun unlocked from the laser beam, (b) optimization of the filter gains and (c) increased data sampling rates. A more rigorous investigation was done on the latter two techniques.

SIMULATION MODEL

Using the vectorial representation of the helicopter-target kinematics shown in figure XI-1, the equations of motion with respect to the X-Y-Z can be derived. Note: The rotation of the x-y-z reference plane attached to the helicopter can be described by

$$\Omega = \omega_x \hat{i} + \omega_y \hat{j} + \omega_z \hat{k}. \qquad \text{(XI-1)}$$

Since there is no rotation about the gun boresight axis,

ω = 0.

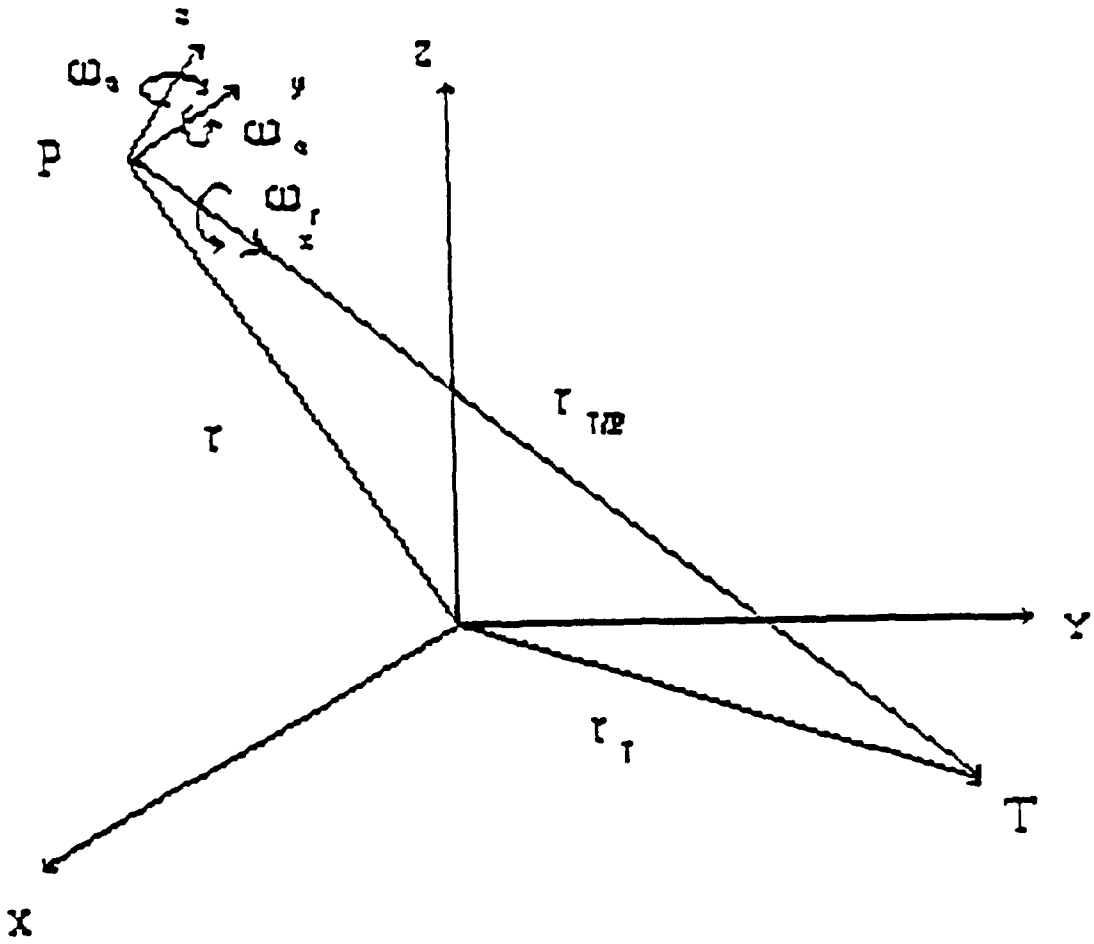


Figure XI-1

Vector Representation of Helicopter-Target Kinematics

It was assumed for this analysis that the target was moving while the platform remained stationary. This has the same results as if the platform was moving with respect to the target, but makes the analysis a bit simpler. The position, velocity and acceleration of the target with respect to the platform are:

$$\begin{aligned}
\vec{r}_T &= r\hat{i}, \\
\vec{v}_T &= \dot{r}\hat{i} + r\omega_a\hat{j} - r\omega_e\hat{k}, \\
\vec{a}_T &= [\ddot{r} - r\omega^2]\hat{i} + [r\dot{\omega}_a + 2\dot{r}\omega_a]\hat{j} - [r\dot{\omega}_e + 2\dot{r}\omega_e]\hat{k},
\end{aligned} \qquad \text{(XI-2)}$$

Where: $\omega^2 = \omega_a^2 + \omega_e^2$

The left hand side of the kinematic acceleration equation above is assumed to be the maneuver noise. The acceleration vector is then used for the state space model of the system kinematics.

There are two stochastic processes at work in the model with respect to system noise. These are (a) noise associated with model uncertainties represented by the maneuver noise and (b) measurement noise. It is assumed that both processes are Gaussian. That is, they have zero mean and are uncorrelated with each other. The units of the maneuver noise are feet/sec^2 and measurement noise for range and angular rates are feet and rad^2/sec^2 respectively.

The kinematic model can be decoupled into three separate sub-models that decribe (a) range and range rate, (b) azimuth rate and (c) elevation rate.

<u>Range and Range Rate</u>

The range state-space model is:

$$\begin{bmatrix} \dot{r} \\ \ddot{r} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \omega^2 & 0 \end{bmatrix} \begin{bmatrix} r \\ \dot{r} \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w_r, \qquad \text{(XI-3)}$$

and

$$r_{meas} = [1 \;\; 0] \begin{bmatrix} r \\ \dot{r} \end{bmatrix} + v_r.$$

The terms w_r and v_r represent the process and measurement noise respectively. Note that the same nomenclature is used for the azimuth and elevation rate models; however, the subscripts are changed to denote the model.

A discrete state-space model as required in a system with discrete measurements can be developed by discretizing the differential equations of motion using the forward difference equation:

$$\dot{x} = \frac{x_{k+1} - x_k}{\Delta T}. \qquad \text{(XI-4)}$$

Thus, the stochastic discrete state-space model for the range and range rate kinematics becomes:

$$\begin{bmatrix} r_{k+1} \\ \dot{r}_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta T_r \\ \Delta T_r \omega^2 & 1 \end{bmatrix} \begin{bmatrix} r_k \\ \dot{r}_k \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w_{r_k}, \tag{XI-5}$$

and

$$r_{meas_k} = [1 \;\; 0] \begin{bmatrix} r_k \\ \dot{r}_k \end{bmatrix} + v_{r_k}. \tag{XI-6}$$

<u>Azimuth and Elevation Rates</u>

The stochastic continuous model for the azimuth rate is:

$$\dot{\omega}_a = -\frac{2\dot{r}}{r}\omega_a + \frac{w_a}{r}, \tag{XI-7}$$

and

$$\omega_{a_{meas}} = \omega_a + v_a. \tag{XI-8}$$

After discretization, the stochastic azimuth rate state-space becomes:

$$\omega_{a_{k+1}} = \left(1 - \frac{2\dot{r}_k \Delta T_a}{r_k}\omega_{a_k}\right) + \frac{\Delta T_a}{r_k} w_{a_k}, \tag{XI-9}$$

and

$$\omega_{a_{meas_k}} = \omega_{a_k} + v_{a_k}. \tag{XI-10}$$

Likewise for the elevation rate, the stochastic continuous state space model is:

$$\dot{\omega}_e = -\frac{2\zeta}{\tau}\omega_e + \frac{w_e}{\tau}, \tag{XI-11}$$

and

$$\omega_{e_{meas}} = \omega_e + v_e. \tag{XI-12}$$

Discretizing the differential equation for the elevation rate:

$$\omega_{e_{k+1}} = \left(1 - \frac{2\zeta\Delta T_e}{\tau_k}\omega_{e_k}\right) + \frac{\Delta T_e}{\tau_k}w_{e_k}, \tag{XI-13}$$

and

$$\omega_{e_{meas_k}} = \omega_{e_k} + v_{e_k}. \tag{XI-14}$$

<u>Kalman Filter</u>

Discrete Kalman filters can now be developed using the above stochastic discrete state-space models. The general form of the Kalman filter is:

$$\bar{X}_{k+1} = \left(\bar{A}_k - \bar{K}_{k+1}\bar{H}_{k+1}\bar{A}_k\right)\bar{X}_k + \bar{K}_{k+1}\bar{Y}_{k+1} \tag{XI-15}$$

Note: the "hatted" terms denotes an estimate, the single bar represents a vector and a double overbar signifies a matrix. For the range and range rate estimators the A_k, H_k and K_k matrices are:

$$\bar{\bar{A}}_k = \begin{bmatrix} 1 & \Delta T_r \\ \Delta T_r \omega^2{}_k & 1 \end{bmatrix}, \quad \bar{\bar{H}}_{k+1} = [1 \;\; 0], \qquad \text{(XI-16)}$$

$$\bar{\bar{K}}_{k+1} = \begin{bmatrix} k_{\rho_{k+1}} \\ k_{v_{k+1}} \end{bmatrix}.$$

The resulting Kalman filters or optimal estimators are:

$$\begin{bmatrix} \hat{\rho}_{k+1} \\ \dot{\hat{\rho}}_{k+1} \end{bmatrix} = \begin{bmatrix} 1 - k_{\rho_{k+1}} & \Delta T_r (1 - k_{\rho_{k+1}}) \\ \Delta T_r \Omega^2{}_k - k_{v_{k+1}} & 1 - \Delta T_r k_{v_{k+1}} \end{bmatrix} \begin{bmatrix} \hat{\rho}_k \\ \dot{\hat{\rho}}_k \end{bmatrix} + \begin{bmatrix} k_{\rho_{k+1}} \\ k_{v_{k+1}} \end{bmatrix} r_{meas_{k+1}}. \qquad \text{(XI-17)}$$

The A_k, H_k and K_k matrices for azimuth and elevation rate estimators are:

$$\bar{\bar{A}}_k = 1 - 2\Delta T_a \frac{\dot{\hat{\rho}}_k}{\hat{\rho}_k}, \quad \bar{\bar{H}}_{k+1} = 1, \quad \bar{\bar{K}}_{k+1} = k_{a_{k+1}}$$

$$\bar{\bar{A}}_k = 1 - 2\Delta T_e \frac{\dot{\hat{\rho}}_k}{\hat{\rho}_k}, \quad \bar{\bar{H}}_{k+1} = 1, \quad \bar{\bar{K}}_{k+1} = k_{e_{k+1}}. \qquad \text{(XI-19)}$$

Using these matrices, the Kalman filters for azimuth and elevation rates become:

$$\Omega_{a_{k+1}} = \left[(1 - k_{a_{k+1}}) \left(1 - 2\Delta T_a \frac{\dot{\hat{\rho}}_k}{\hat{\rho}_k} \right) \right] \Omega_{a_k} - k_{a_{k+1}} \omega_{a_{meas_{k+1}}}, \qquad \text{(XI-20)}$$

and $$\Omega_{e_{k+1}} = \left[\left(1 - k_{e_{k+1}}\right)\left(1 - 2\Delta T_e \frac{\dot{\Omega}_k}{\Omega_k}\right)\right]\Omega_{e_k} + k_{e_{k+1}}\omega_{e_{meas_{k+1}}}, \qquad \text{(XI-21)}$$

The only unknowns in the Kalman filter equations are the filter gains, k_{pr}, k_{vr}, k_{va} and k_{ve}. There are two common approaches to assigning values for these gains. The simplest way is to assign constant values as is the case with the current BSTING algorithms. The other technique is to use <u>a priori</u> and <u>a posteriori</u> state and error covariance equations. These equations update the values of the filter gains at each time increment.

Using constant filter gains works well only if a steady state response is desired. However, if a reliable estimate is desired for the transient response of a filter, then the state and error covariance equations should be solved. This discrete Kalman filter design is often called a predictor-corrector method because the <u>a priori</u> state and error covariance equations are used to predict how the means and variances of the state evolve through time under the influence of the system dynamics. The <u>a posteriori</u> equations correct the states, plus the mean and error covariances based on the updated measurement. The <u>a priori</u> state equation is:

$$\bar{X}_{k+1} = \bar{A}_k \bar{X}_k - \bar{B}_k \bar{U}_k. \qquad \text{(XI-22)}$$

The a priori error covariance equation is:

$$\bar{P}^{-}_{k+1} = \bar{A}_k \bar{P}_k \bar{A}^T_k - \bar{G}_k \bar{Q}_k \bar{G}^T_k \tag{XI-23}$$

Once the a priori state and error covariance are computed, the filter gains and a posteriori states and error covariances can be determined.

$$\bar{X}_{k+1} = \bar{X}^{-}_{k+1} + \bar{K}_{k+1}(\bar{Z}_{k+1} - \bar{H}_{k+1})\bar{P}^{-}_{k+1}, \tag{XI-24}$$

$$\bar{P}_{k+1} = (\bar{I} - \bar{K}_{k+1}\bar{H}_{k+1})\bar{P}^{-}_{k+1}, \tag{XI-25}$$

and

$$\bar{K}_{k+1} = \bar{P}^{-}_{k+1}\bar{H}^T_{k+1}(\bar{H}_{k+1}\bar{P}^{-}_{k+1}\bar{H}^T_{k+1} + \bar{R}_{k+1})^{-1} \tag{XI-26}$$

are the a posteriori state and error covariance and filter gain equations, respectively. The matrices Q_k and R_k are the process noise covariance and measurement noise covariance matrices, respectively. If equations XI-23 and XI-25 are substituted into equation XI-24 and then simplified, the result reduces to equation XI-15 which is the general form of the Kalman filter.

The range a priori error covariance takes the form of a 2 by 2 matrix, the elements of which are given by:

$$
\begin{aligned}
P_{11_{k+1}}^{-} &= P_{11_k} + P_{21_k}\Delta T_r + \Delta T_r(P_{12_k} + P_{22_k}\Delta T_r) \\
P_{12_{k+1}}^{-} &= P_{12_k} + P_{22_k}\Delta T_r + \Delta T_r(P_{11_k} + \Delta T_r P_{21_k})\,\omega_k^2 \\
P_{21_{k+1}}^{-} &= P_{21_k} + \Delta T_r(P_{22_k} + P_{12_k}\Delta T_r\omega_k^2) + P_{11_k}\Delta T_r\omega_k^2 \\
P_{22_{k+1}}^{-} &= P_{22_k} + q_k\Delta T_r^2 + P_{12_k}\Delta T_r\omega_k^2 + \Delta T_r(P_{21_k} + P_{11_k}\Delta T_r\omega_k^2)\,\omega_k^2
\end{aligned}
\tag{XI-27}
$$

where

$$
\bar{P}_k = \begin{bmatrix} P_{11_k} & P_{12_k} \\ P_{21_k} & P_{22_k} \end{bmatrix} \tag{XI-28}
$$

The a posteriori error covariance matrix for the range kinematics is:

$$
\bar{P}_{k+1} = \begin{bmatrix} \dfrac{P_{11_{k+1}}^{-} q_{r_{k+1}}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} & \dfrac{P_{12_{k+1}}^{-} q_{r_{k+1}}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} \\ \dfrac{P_{21_{k+1}}^{-} q_{r_{k+1}}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} & P_{22_{k+1}}^{-} - \dfrac{P_{12_{k+1}}^{-} P_{21_{k+1}}^{-}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} \end{bmatrix} \tag{XI-29}
$$

The filter gains are:

$$
\bar{K}_{k+1} = \begin{bmatrix} \dfrac{P_{11_{k+1}}^{-}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} \\ \dfrac{P_{21_{k+1}}^{-}}{P_{11_{k+1}}^{-} + q_{r_{k+1}}} \end{bmatrix} \tag{XI-30}
$$

The expression for the range and rate state estimators is given in equation XI-16.

The a priori error covariance for the azimuth rate is:

$$p_{k+1}^{-} = \left(1 - 2\frac{\Delta T_s \dot{\rho}_k}{r_k}\right)^2 p_k + \frac{\Delta T_s^2}{r_k^2} q_{a_k}. \qquad \text{(XI-31)}$$

The a posteriori error covariance and filter gain equations are:

$$p_{k+1} = \frac{p_{k+1}^{-} r_{a_{k+1}}}{p_{k+1}^{-} + r_{a_{k+1}}} \qquad \text{(XI-32)}$$

and

$$k_{p_{a_{k+1}}} = \frac{p_{k+1}^{-}}{p_{k+1}^{-} + r_{a_{k+1}}} \qquad \text{(XI-33)}$$

The a priori and a posteriori error covariance and filter gain equations for the elevation are:

$$p_{k+1}^{-} = \left(1 - 2\frac{\Delta T_s \dot{\rho}_k}{r_k}\right)^2 p_k + \frac{\Delta T_s^2}{r_k^2} q_{e_k}, \qquad \text{(XI-34)}$$

$$p_{k+1} = \frac{p_{k+1}^{-} r_{e_{k+1}}}{p_{k+1}^{-} + r_{e_{k+1}}}, \qquad \text{(XI-35)}$$

and

$$k_{p_{e_{k+1}}} = \frac{p_{k+1}^{-}}{p_{k+1}^{-} + r_{e_{k+1}}}. \qquad \text{(XI-36)}$$

COMPUTER SIMULATIONS

Computer programs were written to implement the Kalman filters, error covariances, and filter gains described in the previous section. These were then used to test the validity of the model and analyze the performance of the filter as well as optimize the filter gain coefficients. Two computer programs, Laserpr.m and Kalman.m were written for this simulation. These programs, written in Mathematica, are included at the end of this chapter.

Noise Model and Scenario

In order to simulate the actual measuring and aiming problem, it was necessary to model the "noisy" measurements obtained from the laser rangefinder and azimuth/elevation rate sensors. Again, actual flight test data from the Summer of 1990 tests were used as the basis for the noise models. First, a quadratic least-square expression was fitted to this noisy data. This quadratic expression was then subtracted from the noisy data to obtain a nearly mean generic noise profile. Separate noise profiles were determined for range as well as azimuth and elevation rate measurements. These profiles were then added to the theoretical range and azimuth/elevation rates to obtain the noisy measurement models. Figures XI-2, XI-3 and

XI-4 shows the generic noise profiles that were used in subsequent analyses.

The following flight scenario was used in this simulation:

AIRCRAFT VELOCITY	=	150 knots
ALTITUDE ABOVE TARGET	=	250 feet
RANGE AT CLOSEST APPROACH	=	1000 meters
AZIMUTH RANGE	=	+45 to -45 degrees

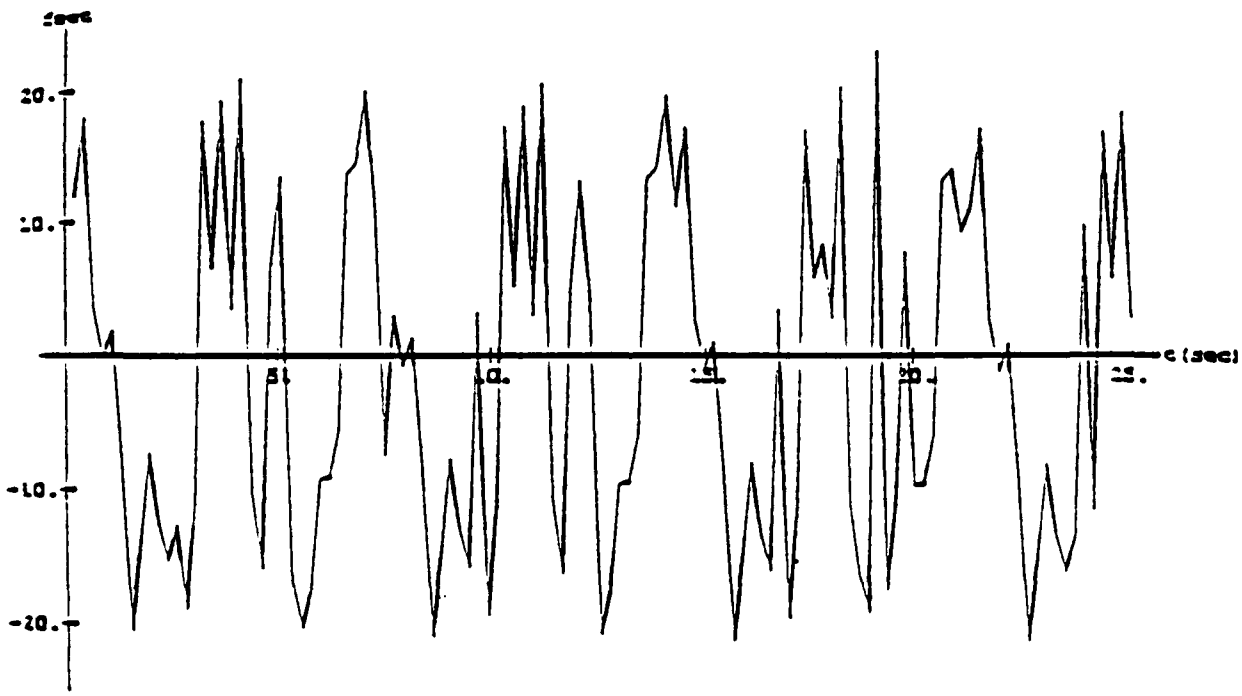


Figure XI-2

Zero-Mean Range Noise Profile

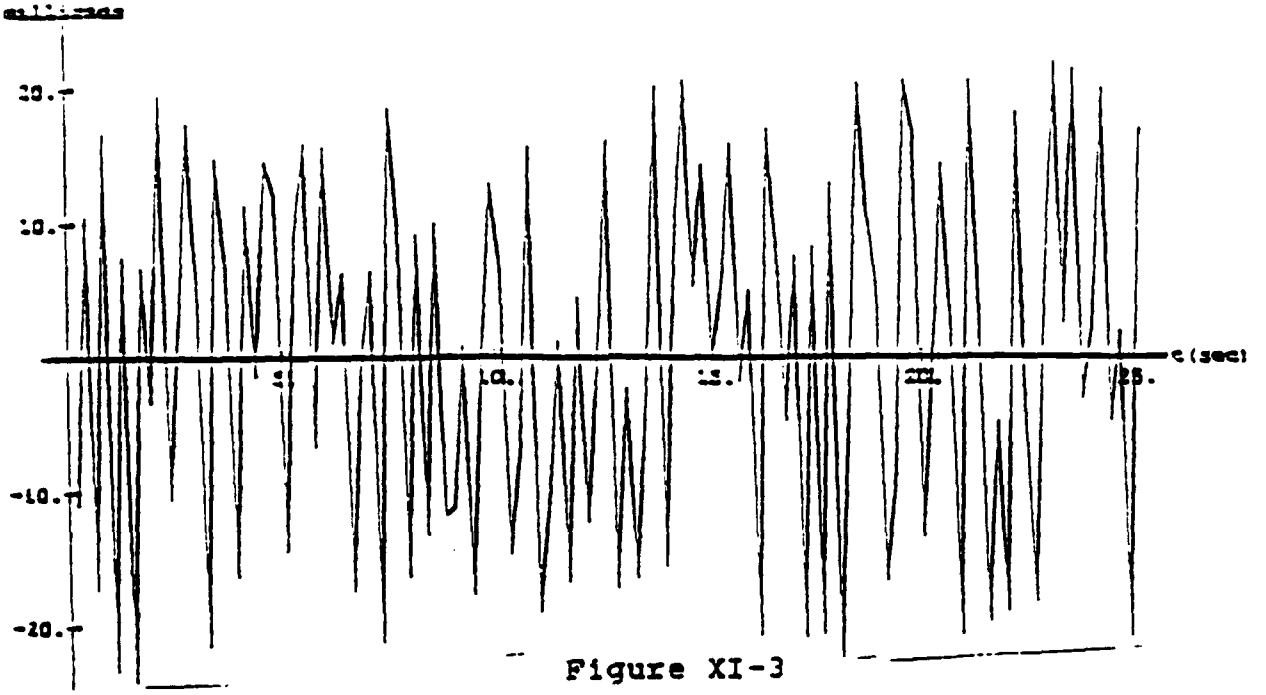


Figure XI-3

Zero-Mean Azimuth Rate Noise Profile

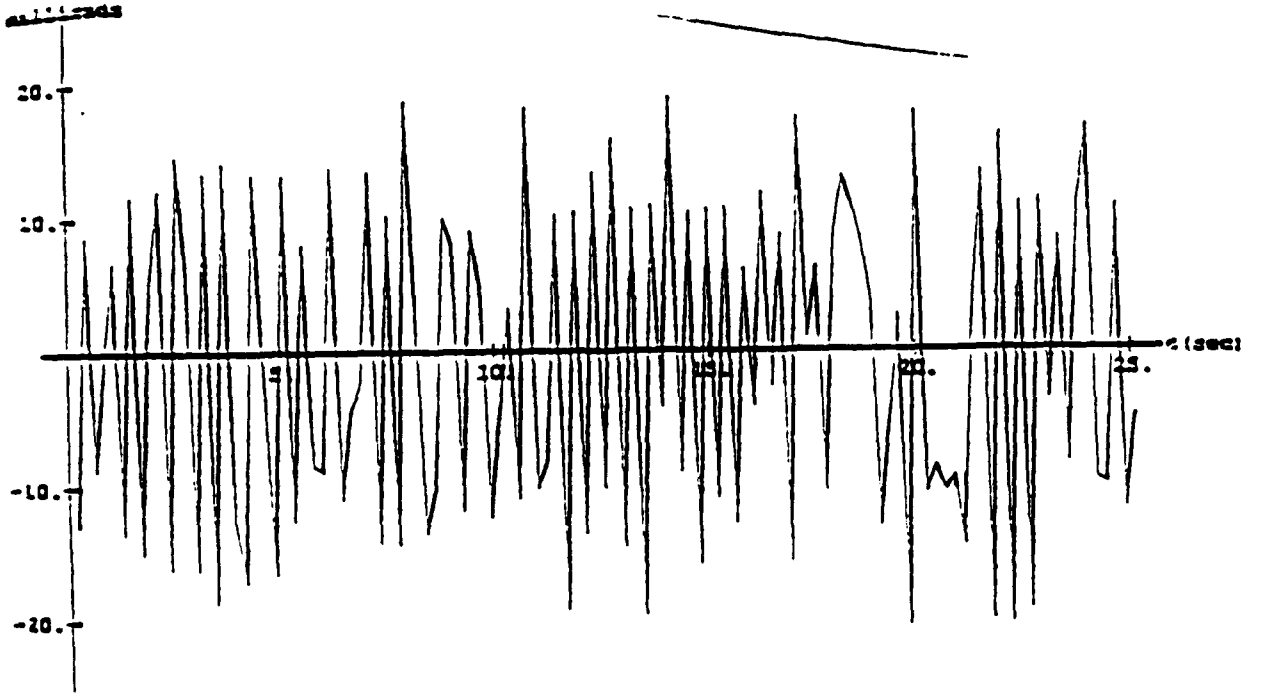


Figure XI-5

Zero-Mean Elevation Rate Noise Profile

Kalman Filter Statistical Parameters

The first step to improve the Kalman filtering was to choose values for the statistical parameters in the respective equations. This was done using a simple sensitivity analysis where the statistical parameters were varied individually over a range of values starting from a baseline set of design parameters. This was a time consuming task because ten parameters had to be evaluated and about seven different values were used for each parameter. Once the simulations were completed the data was reduced and plotted so that effects of changing one variable over the range of values could be analyzed. This sensitivity analysis produced a large matrix of data that was use to choose the optimum parameter values. Optimization was based on the best approximation to the theoretical (unnoisy) azimuth and elevation corrections. The calculated corrections (noisy) had to fall within a +/- 4 milliradian band around the theoretical corrections. The covariance values that gave the best results are shown in table XI-I.

Figure XI-6 shows a comparison between the theoretical (unnoisy) and calculated (noisy) azimuth aiming corrections. The corresponding difference, or error, is shown in Figure XI-7. The results show that the error is outside the +/- 4 milliradians criteria until after 8.5 seconds. After that, the

TABLE XI-I

OPTIMUM COVARIANCE VALUES

$p_{11} = 5000\ ft^2$	$p_{22} = 50\ ft^2/sec^2$
$p_a = 0.03\ rad^2/sec^2$	$p_e = 0.001\ rad^2/sec^2$
$r_r = 10\ ft^2$	$r_a = 0.008\ rad^2/sec^2$
$r_e = 0.01\ rad^2/sec^2$	$q_r = 0.7\ ft^2/sec^2$
$q_a = 0.0006\ rad^2/sec^4$	$q_e = 0.0001\ rad^2/sec^4$

correction stays within error criteria. The corresponding correct -ions in elevation and the elevation aiming errors are shown in figures XI-8 and XI-9, respectively. The elevation correction stays within the +/- 4 milliradian criteria for the entire time.

Figures XI-10 shows how the "adaptive" filter gains varied during the simulation until the they reached the steady state values shown in table XI-II.

TABLE XI-II

STEADY STATE FILTER GAINS

$k_{pr} = 0.15$	$k_{vr} = 0.06$
$k_{pa} = 0.06$	$k_{pe} = 0.03$

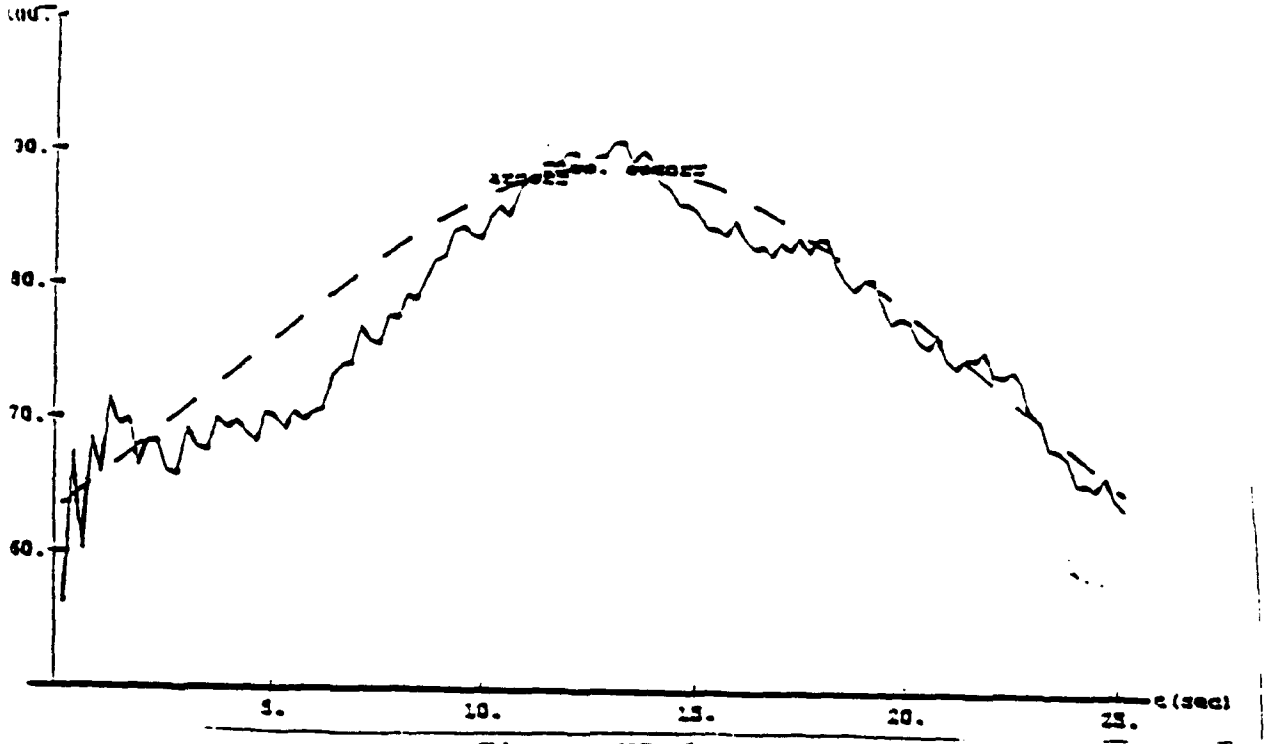


Figure XI-6

Theoretical versus Calculated Azimuth Corrections

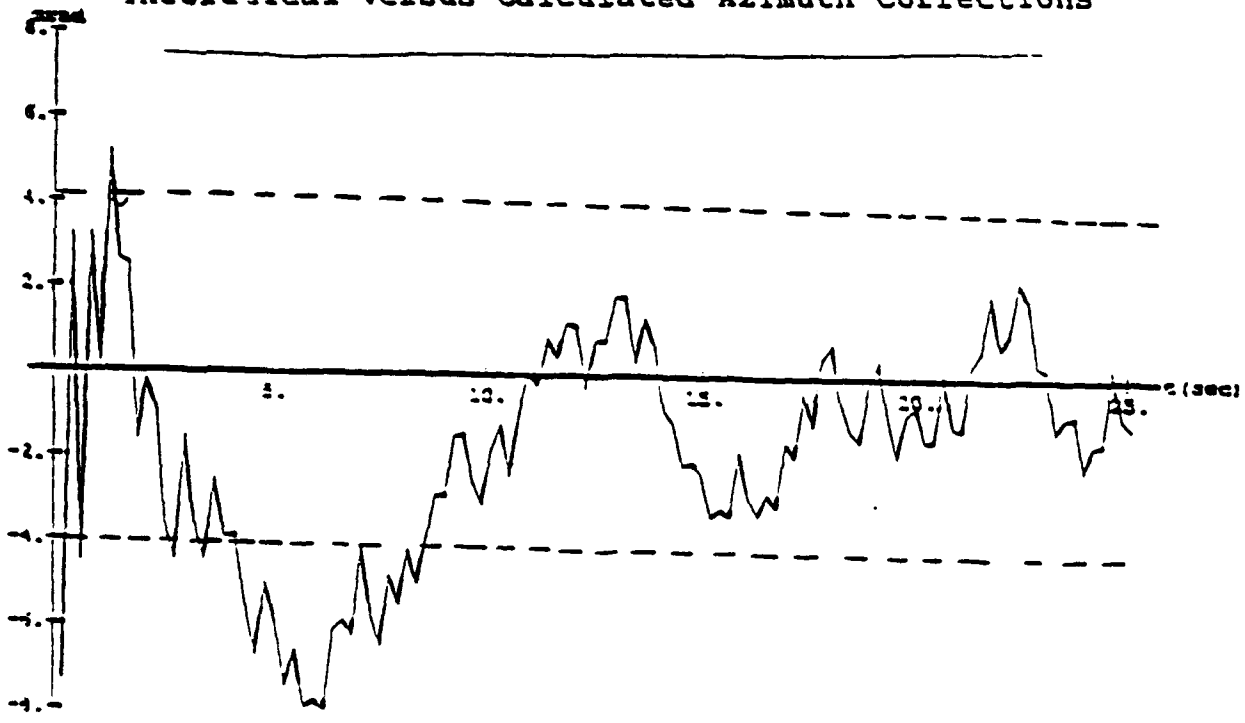


Figure X-7

Error in Azimuth Aiming Correction

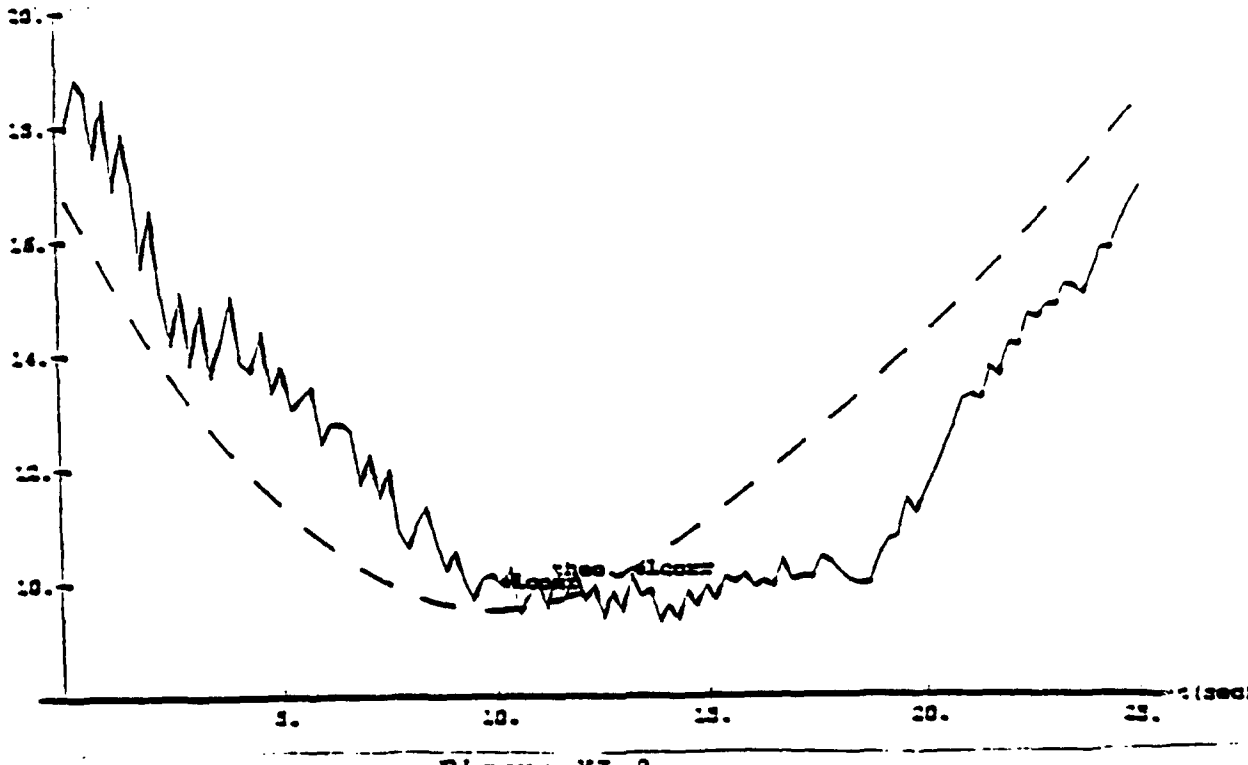


Figure XI-8

Theoretical versus Calculated Elevation Corrections

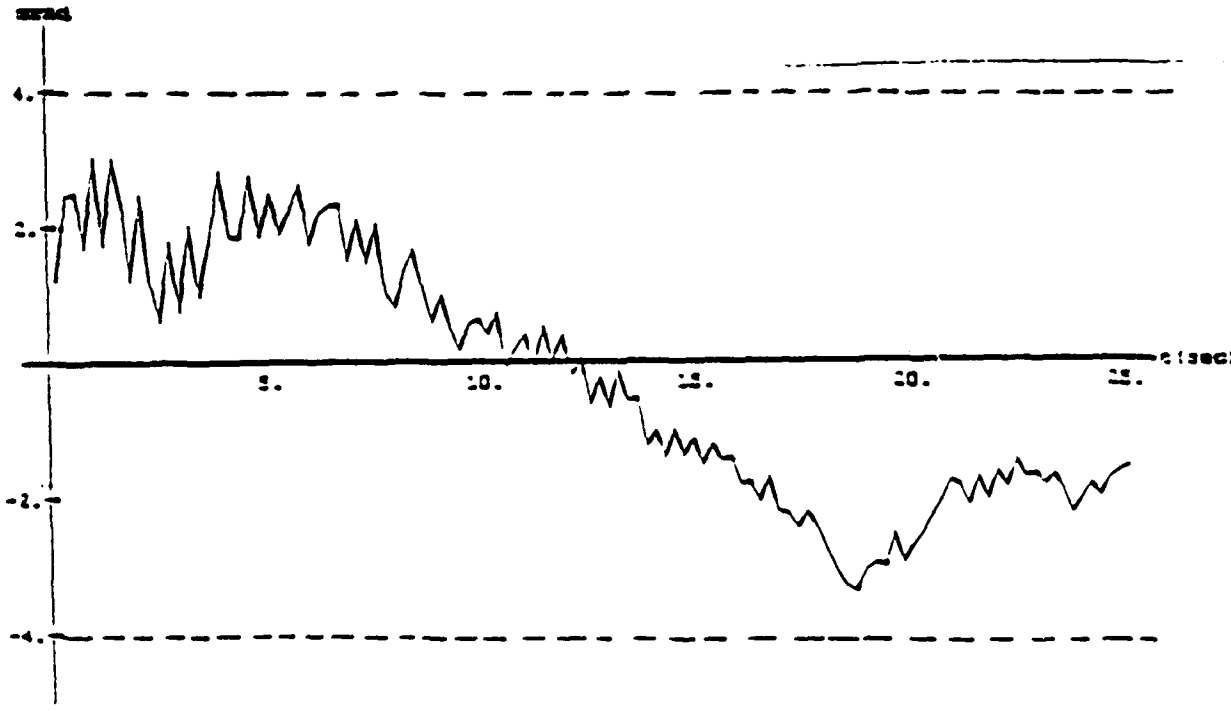


Figure X-9

Error in Elevation Aiming Correction

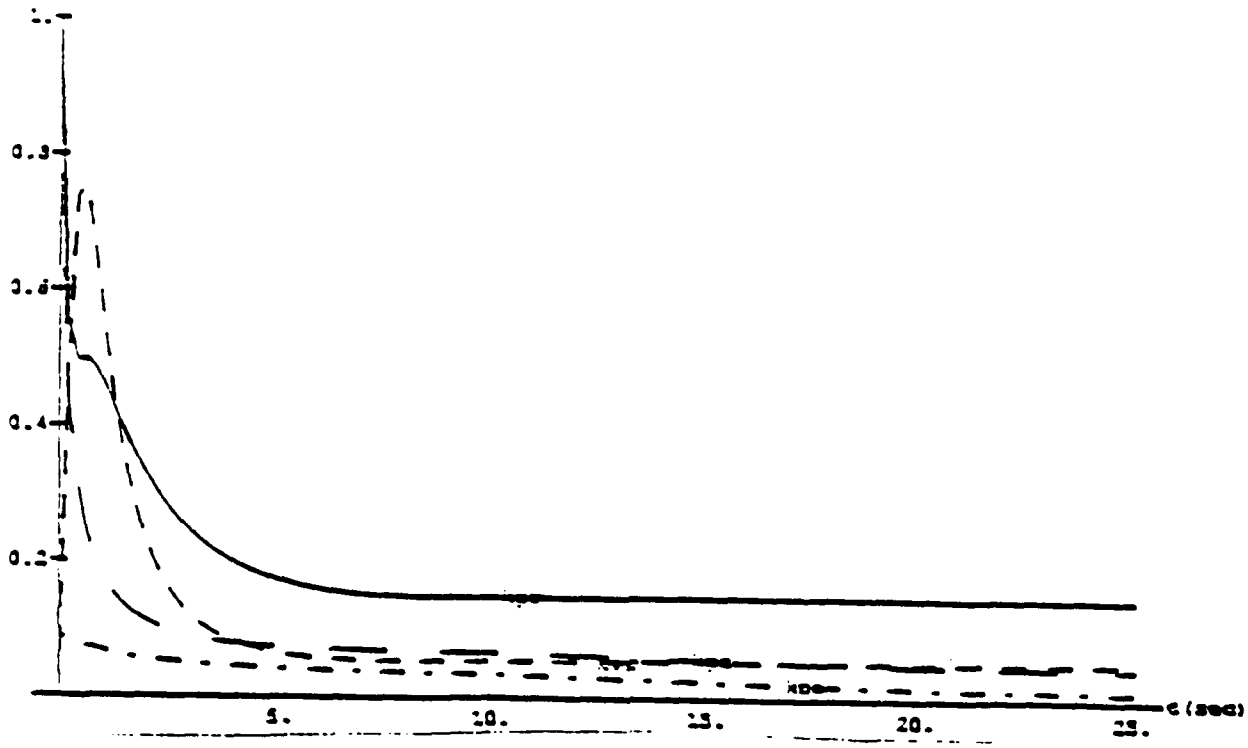


Figure XI-10

Adaptive Filter Gains

Increased Sampling Rates

Next an investigation was made to determine if system response could be improved (i.e. within +/- 4 milliradian error band sooner) by increasing the rate sensor sampling times by factors of two, five or ten times over the current sampling interval of 0.22 seconds. Data acquired at an increased rate was simulated by appropriately compressing the nominal zero mean noise data previously used (e.g. figure XI-3). Sampling intervals of 0.11, 0.044 and 0.022 seconds were investigated.

investigated. In order to obtain sufficiently long noisy files for the shorter sampling rates, the noise files were lengthened by appending duplicate files end-to-end.

The results of this investigation showed a substantial improvement in the response for the azimuth aiming correction and a minor improvement for the elevation aiming correction, which was not a problem in the first place. With a tenfold increase in sampling rate, the +/- 4 milliradians criteria was reached in less than one second compared to about 8.5 second with the 0.22 second sampling rate.

The increased sampling rate data was used with the adaptive filtering. Here a sampling rate of 0.0275 seconds (eightfold increase) was used for azimuth rate data and 0.22 seconds for the elevation rate sampling. Figure XI-11 shows the compressed noise data (0.0275 second sampling rate) added to the theoretical (unnoisy) azimuth rate.

The Kalman filter statistical parameters were again optimized. Again the optimization criteria was the theoretical (unnoisy) aiming corrections. The results of this reoptimization is summarized in table XI-III.

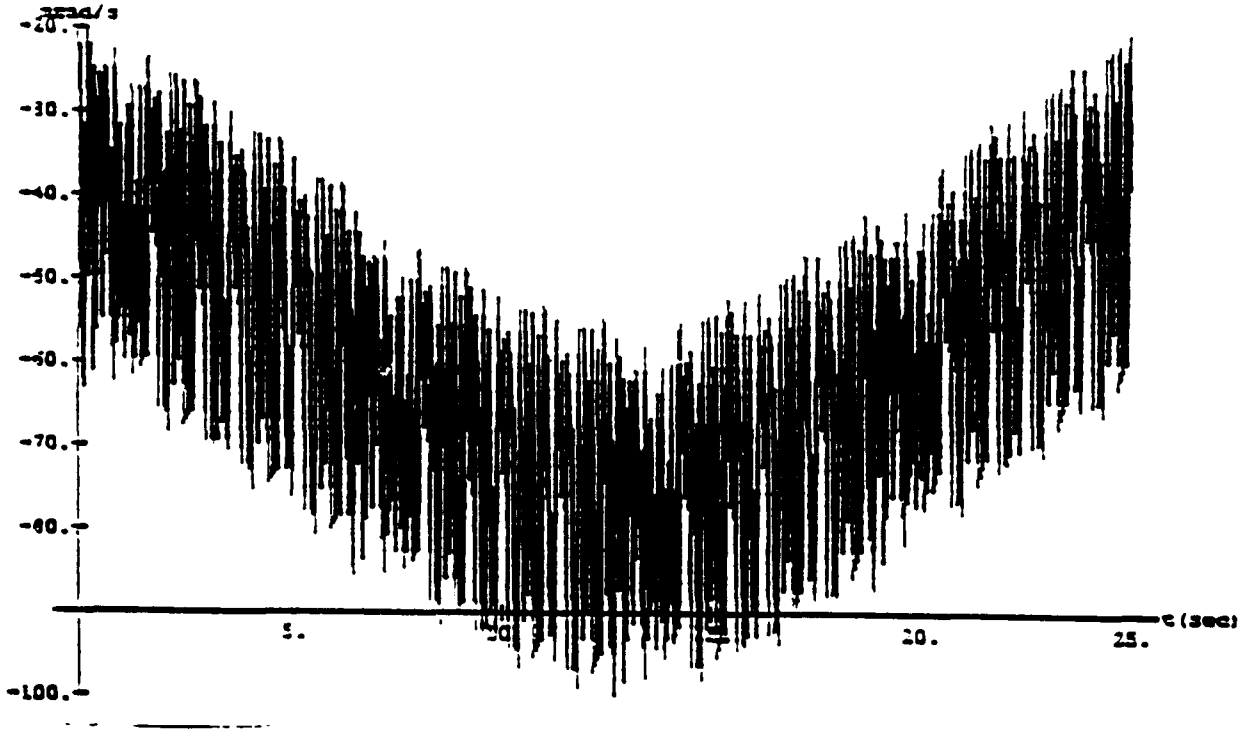


Figure XI-11

Azimuth Rate Sensor Data - 0.0275 second Sampling Rate

TABLE XI-III

REVISED OPTIMUM COVARIANCE VALUES

$p_{11} = 10000\ ft^2$	$p_{22} = 300\ ft^2/sec^2$
$p_a = 0.02\ rad^2/sec^2$	$p_e = 0.003\ rad^2/sec^2$
$r_r = 10\ ft^2$	$r_a = 0.006\ rad^2/sec^2$
$r_e = 0.01\ rad^2/sec^2$	$q_r = 0.7\ ft^2/sec^2$
$q_a = 0.006\ rad^2/sec^4$	$q_e = 0.0005\ rad^2/sec^4$

Using an increased sampling rate results in drastic improvements in system response. Figure XI-10 shows a comparison between the required (theoretical) azimuth aiming correction and the corrections obtained from the noisy rate sensor measurements (0.0275 second sampling interval). The corresponding difference between these corrections is shown in figure XI-11. The difference falls within the +/- 4 milliradian error band within a mere 0.25 seconds. Figures XI-12 and XI-13 shows the corresponding corrections and difference in correction in the elevation plane. Here the error stayed within the required error band for the entire time. Finally, as seen in figure XI-14, while the adaptive filter gains reach steady state values that are similar to those found previously (i.e. figure X-10 and table XI-II), the gains are significantly different during the initial few seconds.

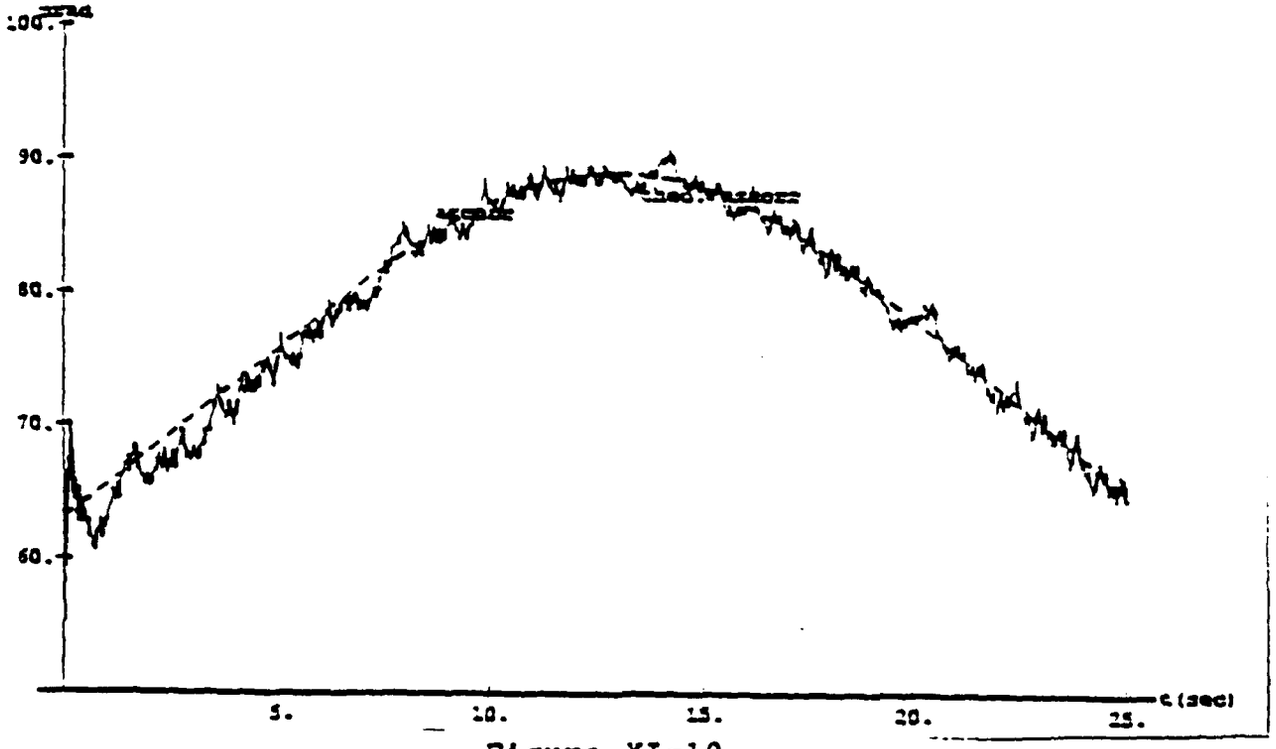


Figure XI-10

Theoretical versus Calculated Azimuth Corrections
(0.0275 second Sampling Rate)

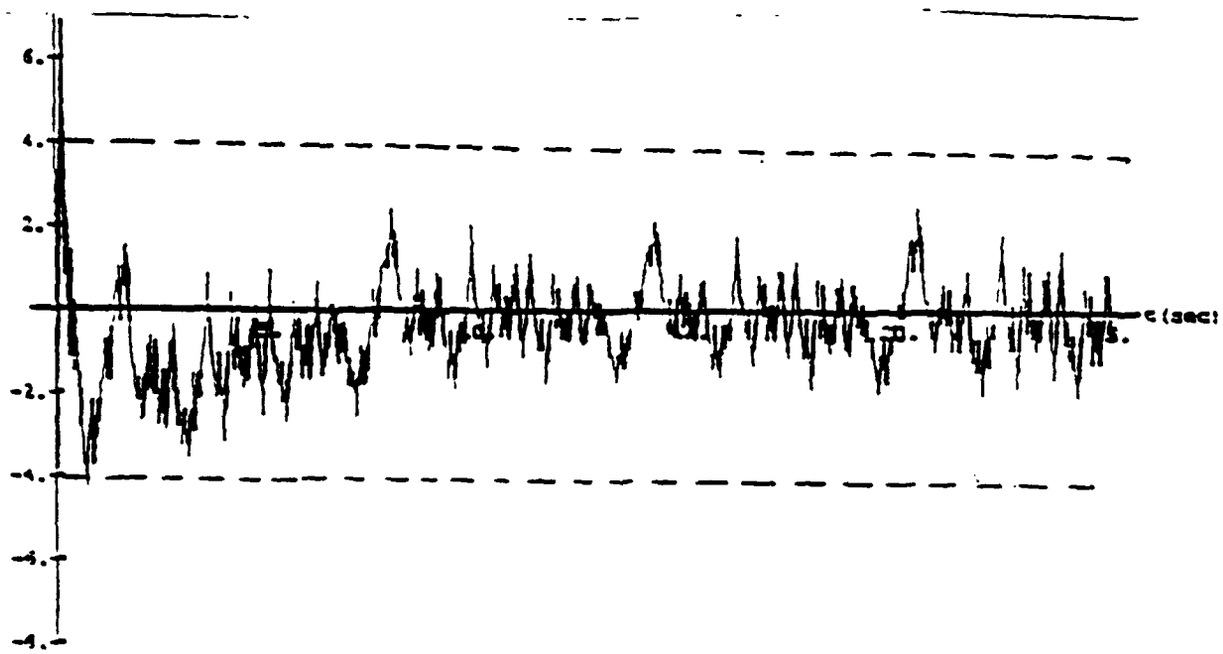


Figure X-11

Error in Azimuth Aiming Correction (0.0275 second Sampling Rate)

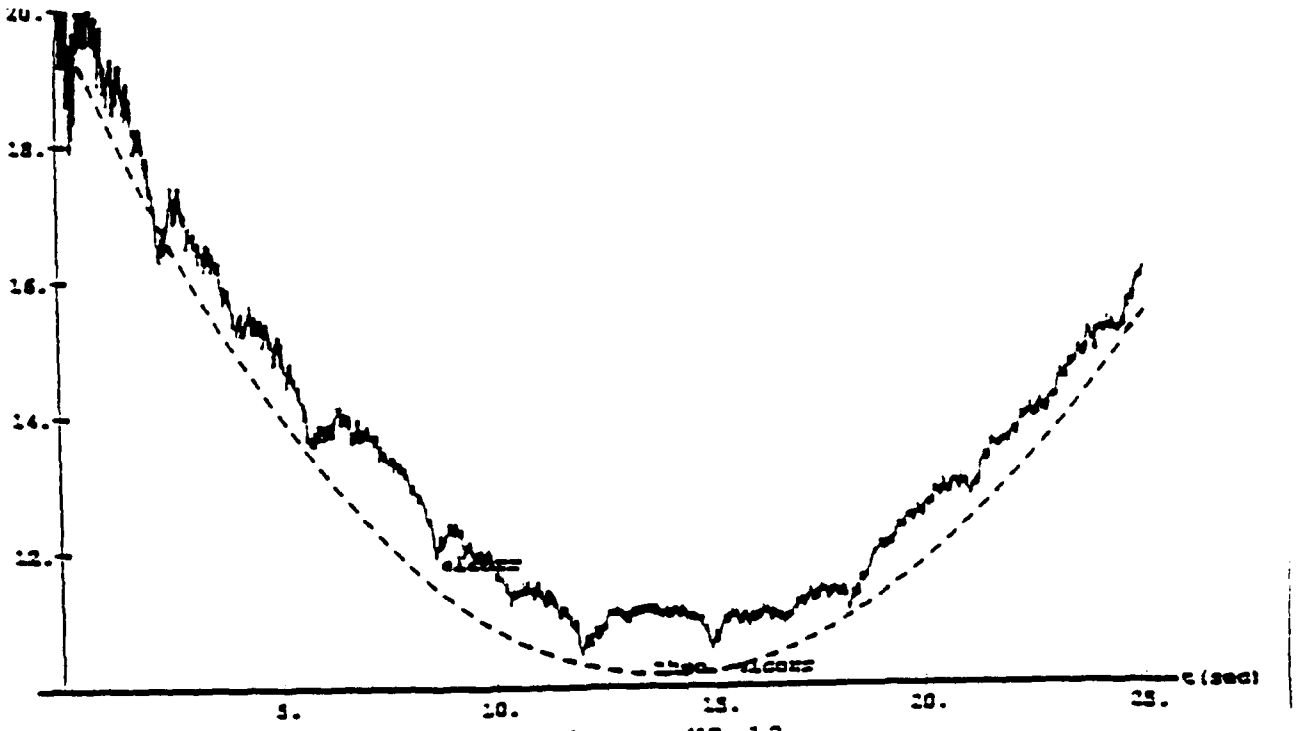


Figure XI-12

Theoretical versus Calculated Elevation Corrections
(0.0275 second Sampling Rate)

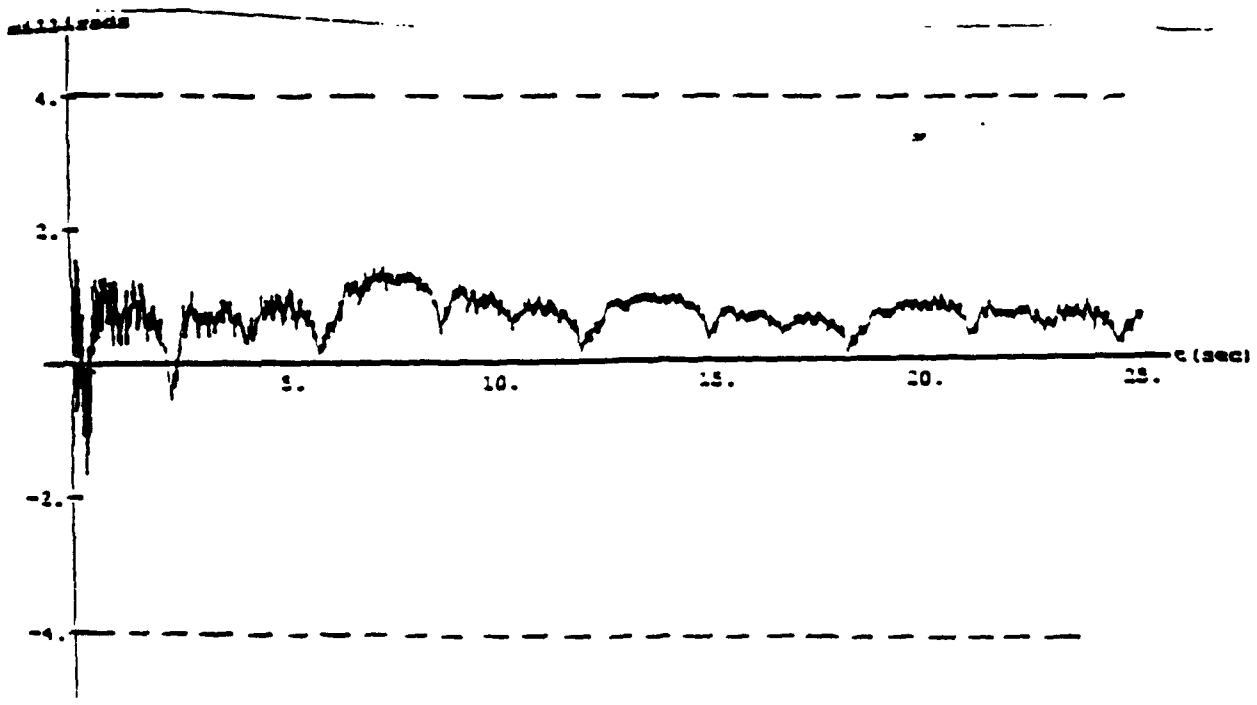


Figure X-13

Error in Elevation Aiming Correction (0.0275 second Sampling Rate)

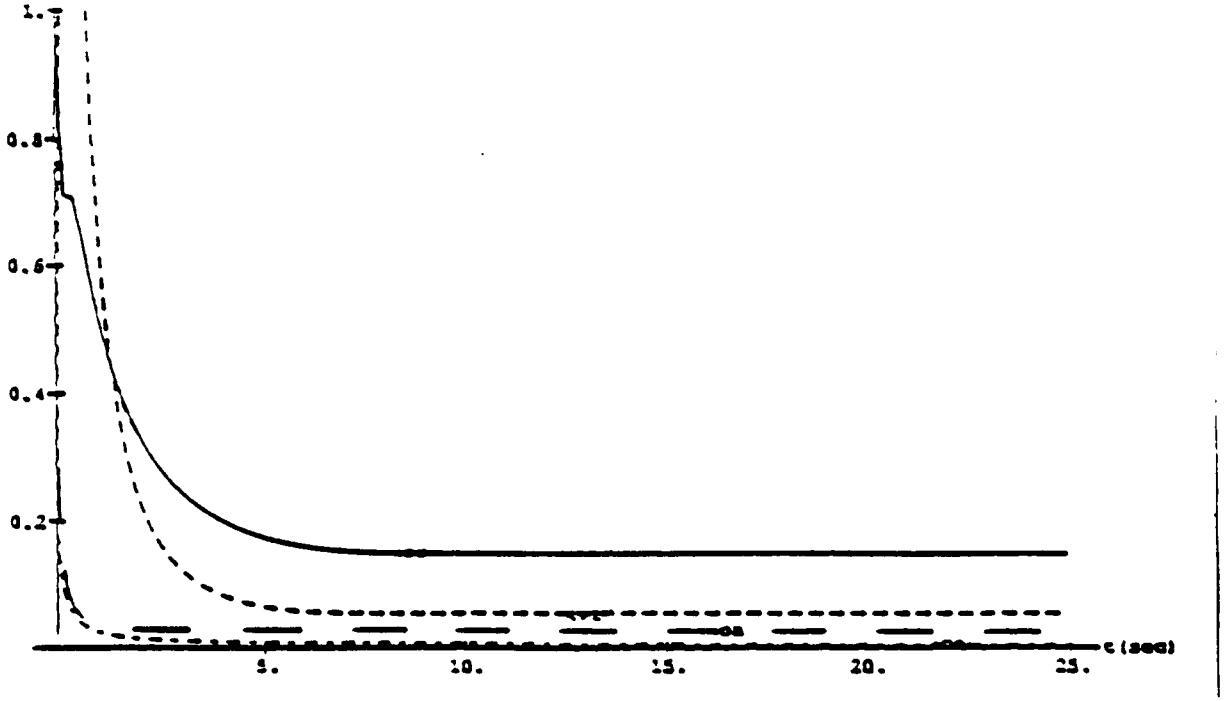


Figure X-14

Revised Adaptive Filter Gains

CHAPTER XII

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

The current BSTING algorithms can be modified to provide a continuous solution so that the gun is always appropriately aimed with respect to the target. This would eliminate the current situation where the aiming correction (difference between gun and laser line-of-sight in azimuth and elevation) is fixed and thus the aiming accuracy degrades as the platform moves with respect to the target.

A correction technique was developed that compensates for the fact that azimuth and elevation rates are measured along the gun boresight rather than the required laser beam axis. Subsequent analysis with "noisy" data simulating an actual gunner aiming from a moving platform showed that measurements made along the gun boresight and laser beam axis provide aiming corrections of essentially equal accuracy.

Furthermore, there was essentially no change in accuracy if the gun and laser were locked together for an initial computation period (i.e. 2 seconds) and then released to allow a continuous tracking and aiming capability.

The current BSTING algorithm for elevation correction due to projectile gravity drop does not account for the elevation angle and uses the "small angle" approximation. An analysis showed that if the platform is over 1000 feet above the target (and a 1 milliradian error is tolerable from gravity drop) then the elevation angle should be measured and included in the computation of the gravity drop correction. The small angle approximation should also not be used.

An alternate fire control system was designed that used external measurement of parameters in lieu of (or in additior to) sensors mounted on the gun only as used in the current BSTING with its gun mounted rate and range sensors. In this system, the actual azimuth and elevation angles between the gun and platform were measured by resolvers. These angles were corrected by the previous aiming correction before being used to calculate the next aiming correction. This technique very closely approximates the use of the more accurate LOS azimuth and elevation angles in calculating aiming corrections.

While this fire control system greatly reduces the sensitivity to azimuth and elevation errors due to gunner pointing difficulties, the system requires accurate measurement of the aircraft velocity. Also the system is more complex than the BSTING system and requires the use of an inertial platform that brings added cost and its own error contributions.

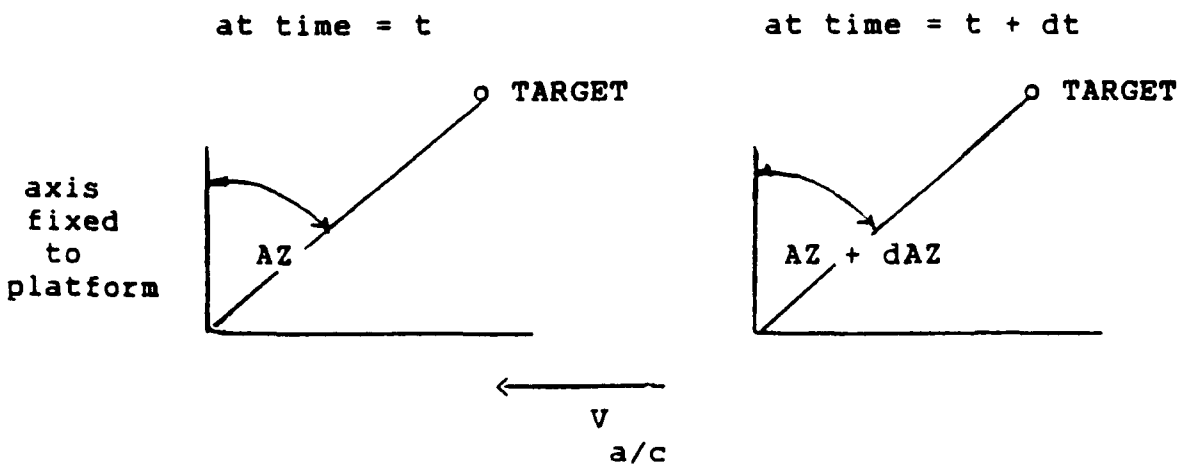
Using constantly updated Kalman filter gains (adaptive Kalman filtering) resulted in a modest improvement in system response (i.e. time before aiming correction errors are within +/- 4 milliradians error band). System response can be substantially improved by increasing the data sampling rate for the sensor measurements, especially in the azimuth plane. For example, when the azimuth rate measurements were simulated at 0.0275 second intervals rather than the 0.22 second sampling rate used in the current BSTING system, the aiming error was within the tolerance band in a mere 0.25 seconds.

RECOMMENDATIONS

While the BSTING system can theoretically be modified to provide continous solutions, an investigation should be done to see the "hardware" effects of actually implementing a continuous solution capability. For example, the dynamics of the servos used to move the laser beam with respect to the gun may result in an additional aiming error or a decreased response.

The increased sampling rate investigates used noisy data simulated simply by taking the BSTING data taken at 0.22 second intervals and reducing the intervals reduced by factors of two, four, eight and ten. A more rigorous analysis should use data from actual sensors that make measurements at the faster rates.

Another idea that might be investigated is to "subtract" out gunner induced noise by using both angular resolvers and rate sensors. This answer would be used to determine platform velocity that could be used in the correction algorithms. For a platform moving with respect to a stationary target:



$$\dot{AZ}_{platform} = dAZ/dt$$

If we know $\dot{AZ}_{platform}$ we can easily compute the platform velocity

i.e. $V_{a/c} = \dot{AZ}_{platform} \times RSL \times \cos(\text{Elevation Angle})$

Now

$$\dot{AZ}_{total} = \dot{AZ}_{platform} + \dot{AZ}_{gun/platform}$$

or

$$\dot{AZ}_{platform} = \dot{AZ}_{total} - \dot{AZ}_{gun/platform}$$

$\dot{AZ}_{total}$ is the total azimuth rate as measured by rate sensors mounted on the gun (e.g. current BSTING type rate sensors)

$\dot{AZ}_{gun/platform}$ is the azimuth rate of the gun with respect to the platform. This could be measured by an angular resolver. (either as angle divided by the time increment or directly as a rate)

Further investigation is needed to see if such a technique will give more accurate results using measurements from actual sensors that could be used in a low cost, gun mounted fire control system.

COMPUTER PROGRAM: Laserpr.m

PURPOSE: Generates theoretical range and azimuth/elevation rate data. Adds generic noise profiles to theoretical data. Computes the theoretical azimuth and elevation aiming corrections that will be used for comparisions of similar corrections based on simulated noisy measurement to establish the performance of the Kalman filter design.

RESULTS: Simulated "noisy" range and azimuth/elevation range measurements used in Kalman.m. Also provides required theoretical aiming corrections used to evaluate system performance.

```
kalman[ ] := Block[

(* objective: This program simulates the discrete kalman filter
designed for the BSTING system. The scenario is a constant velocity
fly by at 150 knots at an altitude of 250 ft., with a minimum range
of 1000 mtr.(3280 ft.). The simulated range, azimuth rate, and
elevation rate data are contained in scen5.dat.                  *)

        { eltheo, aztheo, p11, p12, p21, p22, pa, pe, pb11, pb12,
          pb21, pb22, pba, pbe, kpr, kvr, kpa, kpe, dtr, dtae, rr,
          ra, re, qr, qa, qe, c1, c2, c3, term1, term2, term3,
          rhat, rhatd, azdhat, eldhat, omega2, denom}, {

    azelupdate[ ] := Block[

     (* This procedure calculates the azimuth and elevation
        corrections, due to platform velocity and gravity. The
        calculations also update arrays containing the azimuth and
        elevation rates and azimuth and elevation corrections.
      *)

         { elcorr,  azcorr, vp, wcda, beta, rho, c, tof, gravdrop
         },{

         elcorr = -ArcTan[(rhat*eldhat/2840)/Sqrt[1 -
                  (rhat*eldhat/2840)^2]];
         azc = rhat*azdhat/2840;
         azcorr = -ArcTan[(rhat*azdhat/2840)/Sqrt[1 -
                  (rhat*azdhat/2840)^2]];
         vp = rhatd + 2840*Cos[azcorr]*Cos[elcorr];
         wcda = .000005*rhat^2 - .036938*rhat + 700.429;
         beta = 3.*10^-09*rhat^2 - 4.7*10^-05*rhat + 1.0216;
         rho = .076479;
         c = 1.5*rho/wcda;
         tof = rhat/vp/(1-(c*rhat/4))^2;
         gravdrop = 16.085*(beta*tof)^2/rhat;
         elcorr = elcorr + gravdrop;
         azcorrdata = AppendTo[azcorrdata,{time, azcorr*1000}];
         elcorrdata = AppendTo[elcorrdata,{time, elcorr*1000}]
         }];

(* The next seven procedures generate plots of various parameters
   of interest *)

    plottera[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[

     (* This procedure plots the  azimuth rate estimate data and
        the theoretical azimuth rate data together on one plot. *)

         {style1, style2, plot1, plot2 }, {
         style1 = {Thickness[0.002], PointSize[0.002]};
         style2 = {Dashing[{.01,.01}], Thickness[0.002],
```

```
                PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      grapha = Show[{plot1,plot2,
                 Graphics[ {Text["azdhat", {l1[[350,1]],
                 l1[[350,2]]}], Text["theo. azd", {l2[[70,1]],
                 l2[[70,2]]}] } ]}, PlotLabel -> pl,
                 AxesLabel -> {xl,yl}, PlotRange -> pr ]
      }];

plottere[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[

  (* This procedure plots the  elevation rate estimate data and
     the theoretical elevation rate data together on one plot.
  *)

      {style1, style2, plot1, plot2 }, {
      style1 = {Thickness[0.002], PointSize[0.002]};
      style2 = {Dashing[{.01,.01}], Thickness[0.002],
                 PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      graphe = Show[{plot1,plot2,
                 Graphics[ { Text["eldhat", {l1[[350,1]],
                 l1[[350,2]]}], Text["theo. eld", {l2[[70,1]],
                 l2[[70,2]]}] } ]}, PlotLabel -> pl,
                 AxesLabel -> {xl,yl}, PlotRange -> pr ]
      }];

plotterr[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[

  (* This procedure plots the range estimate data and the
     theoretical range data together on one plot.  *)

      {style1, style2, plot1, plot2 }, {
      style1 = {Thickness[0.002], PointSize[0.002]};
      style2 = {Dashing[{.01,.01}], Thickness[0.002],
                 PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      graphr = Show[{plot1,plot2,
                 Graphics[ { Text["rhat", {l1[[40,1]],
                 l1[[40,2]]}], Text["theo. rng", {l2[[70,1]],
                 l2[[70,2]]}] } ]}, PlotLabel -> pl,
                 AxesLabel -> {xl,yl}, PlotRange -> pr ]
      }];
```

```
plotterd[ l1_, l2_, pl_, xl_, yl_, pr_] := Block{

 (* This procedure plots the range rate estimate data and the
    theoretical range rate data together on one plot. *)

     {style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
              PointSize[0.002]};
     plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
              style1];
     plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
              style2];
     graphrd = Show[{plot1,plot2,
               Graphics[ { Text["rhatd", {l1[[40,1]],
               l1[[40,2]]}], Text["theo. rd", {l2[[70,1]],
               l2[[70,2]]}] } ]}, PlotLabel -> pl,
               AxesLabel -> {xl,yl}, PlotRange -> pr ]
     }];

plotterk[ l1_, l2_, l3_, l4_, pl_, xl_, yl_, pr_] := Block[

(* This procedure overlays the range(kpr), range rate(kvr),
   azimuth rate(kpa), and elevation rate(kpe) filter gain data
 *)

     {style1, style2, style3, style4, plot1, plot2, plot3,
      plot4}, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
              PointSize[0.002]};
     style3 = {Dashing[{.05,.05}], Thickness[0.002],
              PointSize[0.002]};
     style4 = {Dashing[{.005,.01,.01,.01}], Thickness[0.002],
               PointSize[0.002]};
     plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
              style1];
     plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
              style2];
     plot3 := ListPlot[ l3, PlotJoined -> True, PlotStyle ->
              style3];
     plot4 := ListPlot[ l4, PlotJoined -> True, PlotStyle ->
              style4];
     graphk = Show[{plot1,plot2,plot3,plot4,
               Graphics[ { Text["kpr", {l1[[40,1]],
               l1[[40,2]]}], Text["kvr", {l2[[60,1]],
               l2[[60,2]]}], Text["kpa", {l3[[600,1]],
               l3[[600,2]]}], Text["kpe", {l4[[800,1]],
               l4[[800,2]]}] } ]}, PlotLabel -> pl,
               AxesLabel -> {xl,yl}, PlotRange -> pr ]
     }];
```

```
plotterca[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[

(* The azimuth correction based on estimated data and the
   theoretical azimuth correction are plotted on the same
   graph. *)

     {style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
              PointSize[0.002]};
     plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
              style1];
     plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
              style2];
     graphca = Show[{plot1,plot2,
               Graphics[ { Text["azcorr", {l1[[350,1]],
               l1[[350,2]]}], Text["theo. azcorr",
               {l2[[560,1]], l2[[560,2]]}] } ]},
               PlotLabel -> pl, AxesLabel -> {xl,yl},
               PlotRange -> pr ]
     }];

plotterce[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[

(* The elevation correction based on estimated data and the
   theoretical elevation correction are plotted on the same
   graph. *)

   {style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
              PointSize[0.002]};
     plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
              style1];
     plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
              style2];
     graphce = Show[{plot1,plot2,
               Graphics[ { Text["elcorr", {l1[[350,1]],
               l1[[350,2]]}], Text["theo. elcorr",
               {l2[[560,1]], l2[[560,2]]}] } ]},
               PlotLabel -> pl, AxesLabel -> {xl,yl},
               PlotRange -> pr ]
     }];

(* The main body of the program begins here.

  Define the necessary data arrays.                *)

   rhatdata = {};      (* range estimate              *)
   rhatddata = {};     (* range rate estimate         *)
   azdhatdata = {};    (* azimuth rate estimate       *)
   eldhatdata = {};    (* elevation rate estimate     *)
```

```
kprdata = {};        (* range filter gain              *)
kvrdata = {};        (* range rate filter gain         *)
kpadata = {};        (* azimuth rate filter gain       *)
kpedata = {};        (* elevation rate filter gain     *)
azcorrdata = {};     (* azimuth correction             *)
elcorrdata = {};     (* elevation correction           *)

(* Open and read into an array the simulated noisy range, azimuth
   rate, and elevation rate measurements. *)

data = <<scen5.dat;

(* initialize state covariance matricies. The noise is assumed to
   be gaussian so the range states noises are uncorrelated. Thus
   p12 and p21 are zero.    *)

p12 = 0;
p21 = 0;
Print["Enter the initial state covariances."];
p11 = Input["p11 = ? "];
p22 = Input["p22 = ? "];
pa = Input["pa = ? "];
pe = Input["pe = ? "];
Print[];

(* user initializes measurement error covariances.    *)
Print["Enter the measurement error covariances."];
rr = Input["rr = ? "];
ra = Input["ra = ? "];
re = Input["re = ? "];
Print[];

(* user initializes plant error covariances.    *)
Print["Enter the plant error covariances."];
qr = Input["qr = ? "];
qa = Input["qa = ? "];
qe = Input["qe = ? "];

(* initialize remaining variables          *)
time = 0;
dtr = .22;      (* dtr is the range sampling rate in seconds.*)
dtae = .0275;   (* dtae is the sampling rate of az. and el.
                   rate sensors. *)
rhat = data[[1,4]];
rhatd = 0;
azdhat = 0;
eldhat = 0;

(* the simulation evolves at two different rates. The range
   filter has a sampling rate of .22s, while the azimuth rate
   and elevation rate filters have sampling rates of .0275s.
 *)
```

```
Do[(
     time = data[[i,1]];

  (* update az. and el. rate a priori error covariances    *)
     pba = (1 - 2*dtae*rhatd/rhat)^2*pa + dtae^2*qa;
     pbe = (1 - 2*dtae*rhatd/rhat)^2*pe + dtae^2*qe;

  (* update az. and el. rate filter gains    *)
     kpa = pba/(pba + ra );
     kpe = pbe/(pbe + re );

  (*  update az. and el. rate a posteriori error covariances *)
     pa = (1 - kpa)*pba;
     pe = (1 - kpe)*pbe;

   (* calculate current az. and el. rate estimates    *)
     term1 = 1 - kpe;
     term2 = 1 - kpa;
     eldhat = term1*(1 - 2*rhatd*dtae/rhat)*eldhat +
                kpe*data[[i,3]]/1000;
     azdhat = term2*(1 - 2*rhatd*dtae/rhat)*azdhat +
               kpa*data[[i,2]]/1000;

  (* calculate the range estimate every eighth time step   *)
     If[ Mod[(i-1), 8] == 0,
        omega2 = azdhat^2 + eldhat^2;

  (* update range a priori error covariances                *)
        pb11 = p11 + dtr*(p12 + p21 + p22*dtr);
        pb12 = p12 + dtr*(p22 + (p11 + dtr*p21)*omega2);
        pb21 = p21 + dtr*(p22 + (p11 + dtr*p12)*omega2);
        pb22 = p22 + qr*dtr^2 + omega2*dtr*(p21 + p12 +
                dtr*omega2*p11);

   (* update range filter gains           *)
        denom = pb11 + rr;
        kpr = pb11/denom;
        kvr = pb21/denom;

   (* update range a posteriori error covariances   *)
        p11 = pb11*rr/denom;
        p12 = pb12*rr/denom;
        p21 = pb21*rr/denom;
        p22 = pb22 - pb21*pb12/denom;

     (* calculate current range and range rate estimates *)
        c1 = 1 - kpr;
        c2 = dtr*c1;
        c3 = 1 - kvr*dtr;
        rhat = c1*rhat + c2*rhatd + kpr*data[[i,4]];
        term3 = omega2*dtr;
        rhatd = rhat*(term3 - kvr) + rhatd*c3 -
```

```
                    data[[i,4]]*kvr;

(* add current range estimates, range rate estimates, and filter
   gains to time history arrays  *)

            kprdata = AppendTo[ kprdata, {time,kpr}];
            kvrdata = AppendTo[ kvrdata, {time,kvr}];
            rhatdata = AppendTo[ rhatdata, {time,rhat}];
            rhatddata = AppendTo[ rhatddata, {time,rhatd}] ];

(* add current az. and el. rate estimates and filter gains to
   time history arrays *)
        azdhatdata = AppendTo[ azdhatdata, {time,azdhat*1000}];
        eldhatdata = AppendTo[ eldhatdata, {time,eldhat*1000}];
        kpadata = AppendTo[ kpadata, {time,kpa}];
        kpedata = AppendTo[ kpedata, {time,kpe}];

 (* calculate corrected azimuth and elevation angles *)
        azelupdate[]},
      {i, 1, Length[data]}];

  (* generate plots of time histories
     These arrays were generated by the program laserpr.m. They
     contain theoretical values.   *)

      aztheo = <<aztheo.dat;      (* theo. azimuth rate          *)
      eltheo = <<eltheo.dat;      (* theo. elevation rate        *)
      rngtheo = <<rngtheo.dat;    (* theo. range                 *)
      rngdtheo = <<rngdtheo.dat;  (* theo. range rate            *)
      azcrtheo = <<azc5.dat;      (* theo. azimuth correction    *)
      elcrtheo = <<elc5.dat;      (* theo. elevation correction *)
      azcorrdata = Rest[azcorrdata];
      elcorrdata = Rest[elcorrdata];
      azcrdiff = Table[ {azcorrdata[[i,1]], azcorrdata[[i,2]] -
                 azcrtheo[[i,2]]}, {i,1,Length[azcorrdata]}];
      elcrdiff = Table[ {elcorrdata[[i,1]], elcorrdata[[i,2]] -
                 elcrtheo[[i,2]]}, {i,1,Length[elcorrdata]}];
      plotterca[azcorrdata,azcrtheo, "azcorr & theo. azcorr for
                scenario:", "t(sec)", "mrad", {50,100}];
      plotterda = ListPlot[azcrdiff, PlotJoined -> True,
                  PlotRange -> {-5,5}, AxesLabel ->
                  {"t(sec)","mrad"}, PlotLabel -> "difference
                  between azcorr and theo. azcorr"],
      plotterce[elcorrdata,elcrtheo, "elcorr & theo elcorr for
                scenario:", "t(sec)", "mrad/s", {10,20}];
      plotterde = ListPlot[elcrdiff, PlotJoined -> True,
                  PlotRange -> {-5,5}, AxesLabel ->
                  {"t(sec)","mrad"}, PlotLabel -> "difference
                  between elcorr and theo. elcorr"],
      plottera[azdhatdata,aztheo, "azdhat & theo. az. rate for
               scenario:", "t(sec)", "mrad/s", {-80,-30}];
      plottere[eldhatdata,eltheo, "eldhat & theo el. rate for
```

```
            scenario:", "t(sec)", "mrad/s", {-12,4}];
  plotterr[rhatdata,rngtheo, "rdhat & theo range for
            scenario:", "t(sec)", "feet", {3000,5000}];
  plotterd[rhatddata,rngdtheo, "rhatd & theo range rate for
            scenario:", "t(sec)", "ft/s", {-200,200}];
  plotterk[kprdata,kvrdata,kpadata,kpedata,
           "filter gains for scenario:", "t(sec)",  " ",
           {0,1}];
}]
```

```
laserpr[ ] := Block[  {
                        vac, rmin, h, rsl, elevation, rsldot, azdot,
                        eldot, elcorr, velproj, wcda, beta, rho, c,
                        tof, azcorr, azdot, rngnoise, elnoise, noise
                        }, {

   vac = 1.6889*Input["Aircraft Velocity(knots) = "];
   rmin = 3.28*Input["Minimum Range(meters) = "];
   h = Input["Altitude(feet) = "];
   azimuth = 0;
   grav = 32.17;
   time = 0.0;
   i = 1;
   ell = elevation;
   eldotdata = {};
   azdotdata = {};
   rngdata = {};
   rngddata = {};
   elcordata = {};
   azcordata = {};
   While[azimuth > -.75049,
         {rbase = rmin - vac*time;
         rsl = Sqrt[h^2 + rmin^2 + rbase^2];
         azimuth = ArcTan[rbase/rmin];
         elevation = ArcTan[h/Sqrt[rmin^2 + rbase^2]];
         rsldot = -vac*rbase/rsl;
         azdot = -vac*rmin/rsl^2;
         eldot = (elevation - ell)/.22;
         ell = elevation;
         elcorr = rsl*eldot/2840;
         elcorr = -ArcTan[elcorr/Sqrt[1 - elcorr^2]];
         azcorr = rsl*azdot/(2840*Cos[elcorr]);
         azcorr = -ArcTan[azcorr/Sqrt[1 - azcorr^2]];
         velproj = rsldot + 2840*Cos[elcorr]*Cos[azcorr];
         wcda = .000005*rsl^2 - .036938*rsl + 700.4299;
         beta = 3.*10^-09*rsl^2 - 4.7*10^-05*rsl + 1.0216;
         rho = .076475;
         c = 1.5*rho/wcda;
         tof = rsl/velproj/(1-c*rsl/4)^2;
         elgravdrop = (grav/2)*((beta*tof)^2)/rsl;
         elcorr = elcorr + elgravdrop;
         If[ i == 1 || Mod[(i-1), 8] == 0, range = rsl];
         azdotdata = AppendTo[azdotdata,{time, azdot*1000}];
         eldotdata = AppendTo[eldotdata,{time, eldot*1000}];
         rngdata = AppendTo[rngdata,{time, range}];
         rngddata = AppendTo[rngddata,{time,rsldot}];
         azcordata = AppendTo[azcordata,{time, azcorr*1000}];
         elcordata = AppendTo[elcordata,{time, elcorr*1000}];
         Print[ i,"  ", time, "  ", rsl, "  ", range];
         i += 1;
         time += .0275}];
   Rest[azcordata];
```

```
   Rest[elcordata];
   elnoise = <<nosel4.dat;
   aznoise = <<nosaz4.dat;
   rngnoise = <<nosrng4.dat;
   nazdotdata = Table[ {azdotdata[[i,1]], azdotdata[[i,2]] +
                        aznoise[[i,2]]}, {i,Length[azdotdata]}];
   neldotdata = Table[ {eldotdata[[i,1]], eldotdata[[i,2]] +
                       elnoise[[i,2]]}, {i,Length[eldotdata]}];
   nrngdata = Table[ {rngdata[[i,1]], rngdata[[i,2]] +
                     rngnoise[[i,2]]}, {i,Length[rngdata]}];
   Print["Theoretical data for azimuth rate, elevation rate,
azimuth"];
   Print["correction, elevation correction, and range are in the
arrays"];
   Print["azdotdata, eldotdata, azcordata, elcordata, and rngdata,
having"];
   Print["the dimensions milliradians, milliraadians per second,
and feet."];
   Print[" "];
   Print["Noise corrupted data for azimuth rate, elevation rate,
and range"];
   Print["are in the arrarys nazdotdata, neldotdata, nrngdata."];
  } ]
```

```
filt3[ ] := Block[ { }, {

     azelupdate[ ] := Block[ { }, {
      (* this procedure calculates the azimuth and elevation
         corrections. It also updates arrays containing the
         azimuth and elevation rates and azimuth and elevation
         corrections.  *)

          elc = rhat*eldhat/2840;
          elcorr = -ArcTan[ elc/Sqrt[ 1 - elc^2 ] ];
          azc = rhat*azdhat/2840;
          azcorr = -ArcTan[ azc/Sqrt[ 1 - azc^2 ] ];
          vp = rhatd + 2840*Cos[azcorr]*Cos[elcorr];
          wcda = .000005*rhat^2 - .036938*rhat + 700.429;
          beta = 3.*10^-09*rhat^2 - 4.7*10^-05*rhat + 1.0216;
          rho = .076479;
          c = 1.5*rho/wcda;
          tof = rhat/vp/(1-(c*rhat/4))^2;
          gravdrop = 16.085*(beta*tof)^2/rhat;
          elcorr = elcorr + gravdrop;
          azcorrdata = AppendTo[azcorrdata,{time, azcorr*1000}];
         elcorrdata = AppendTo[elcorrdata,{time, elcorr*1000}]
      }];

     plottera[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
          style1, style2, plot1, plot2 }, {
          style1 = {Thickness[0.002], PointSize[0.002]};
          style2 = {Dashing[{.01,.01}], Thickness[0.002],
                    PointSize[0.002]};
          plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                   style1];
          plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                   style2];
         grapha = Show[{plot1,plot2, Graphics[ { Text["azdhat",
                  {l1[[50,1]], l1[[50,2]]}], Text["theo. azd",
                  {l2[[60,1]], l2[[60,2]]}] } ]}, PlotLabel -> pl,
                   AxesLabel -> {xl,yl},PlotRange -> pr ]
          }];

     plottere[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
          style1, style2, plot1, plot2 }, {
          style1 = {Thickness[0.002], PointSize[0.002]};
          style2 = {Dashing[{.01,.01}], Thickness[0.002],
                   PointSize[0.002]};
          plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                   style1];
          plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                   style2];
          graphe = Show[{plot1, plot2, Graphics[ { Text[ "eldhat",
                   {l1[[50,1]], l1[[50,2]]}], Text["theo. eld",
                   {l2[[60,1]], l2[[60,2]] } ] } ]},
                   PlotLabel -> pl, AxesLabel -> {xl,yl},
```

```
                PlotRange -> pr ]
     }];

plotterr[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
     style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
               PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      graphr = Show[ {plot1, plot2, Graphics[ {Text["rhat",
                {l1[[50,1]], l1[[50,2]]}], Text[ "theo. rng",
                {l2[[60,1]], l2[[60,2]]}] } ]},
                PlotLabel -> pl, AxesLabel -> {xl,yl},
                PlotRange -> pr ]
     }];

plotterd[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
     style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
               PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      graphrd = Show[{plot1, plot2, Graphics[ { Text["rhatd",
                 {l1[[50,1]], l1[[50,2]]}], Text["theo. rd",
                 {l2[[60,1]], l2[[60,2]]}] } ]},
                 PlotLabel -> pl, AxesLabel -> {xl,yl},
                 PlotRange -> pr ]
     }];

plotterca[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
     style1, style2, plot1, plot2 }, {
     style1 = {Thickness[0.002], PointSize[0.002]};
     style2 = {Dashing[{.01,.01}], Thickness[0.002],
                PointSize[0.002]};
      plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                style1];
      plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                style2];
      graphca = Show[{plot1, plot2, Graphics[ { Text["azcorr",
                 {l1[[50,1]], l1[[50,2]]}], Text["theo.
                 azcorr", {l2[[60,1]],l2[[60,2]]}] } ]},
                 PlotLabel -> pl, AxesLabel -> {xl,yl},
                 PlotRange -> pr ]
     }];

plotterce[ l1_, l2_, pl_, xl_, yl_, pr_] := Block[ {
```

```
        style1, style2, plot1, plot2 }, {
        style1 = {Thickness[0.002], PointSize[0.002]};
        style2 = {Dashing[{.01,.01}], Thickness[0.002],
                 PointSize[0.002]};
         plot1 := ListPlot[ l1, PlotJoined -> True, PlotStyle ->
                  style1];
         plot2 := ListPlot[ l2, PlotJoined -> True, PlotStyle ->
                  style2];
         graphce = Show[{plot1,plot2,
                   Graphics[ { Text["elcorr", {l1[[50,1]],
                   l1[[50,2]]}], Text["theo. elcorr",
                   {l2[[60,1]], l2[[60,2]]}] } ]},
                   PlotLabel -> pl, AxesLabel -> {xl,yl},
                   PlotRange -> pr ]
        }];

    rhatdata = {};
    rhatddata = {};
    azdhatdata = { };
    eldhatdata = { };
    azcorrdata = {};
    elcorrdata = {};
    rngnoise = <<rngnoise.dat;
    aznoise = <<aznoise.dat;
    elnoise = <<elnoise.dat;
    data = Table[ {rngnoise[[i,1]], aznoise[[i,2]],
           elnoise[[i,2]], rngnoise[[i,2]]},
           {i,Length[rngnoise]}];
    rngnoise = .;
    aznoise = .;
    elnoise = .;
    kpa = Input["KPA = ?"];
    kpe = Input["KPE = ?"];
    kpr = Input["KPR = ?"];
    kvr = Input["KVR = ?"];
    dtr = .22;

(*  first sample

    time = data[[i,1]]
    gxhat = data[[i,2]]
    gyhat = data[[i,3]]
    senrng = data[[i,4]]      *)

    time = data[[1,1]];
    rhat = data[[1,4]];
    azdhat = data[[1,2]]/1000;
    eldhat = data[[1,3]]/1000;
    azdhatdata = AppendTo[ azdhatdata, {time,azdhat*1000}];
    eldhatdata = AppendTo[ eldhatdata, {time,eldhat*1000}];
    rhatdata = AppendTo[ rhatdata, {time,rhat}];
    rhatddata = AppendTo[ rhatddata, {time,rhatd}];
```

```
     azelupdate[ ];

(*    second sample              *)

     time = data[[2,1]];
     rhat = .5*(data[[2,4]] + rhat);
     rhatd = (data[[2,4]] - rhat)/dtr;
     azdhat = (azdhat + data[[2,2]]/1000)/2;
     eldhat = (eldhat + data[[2,3]]/1000)/2;
     azdhatdata = AppendTo[ azdhatdata, {time,azdhat*1000}];
     eldhatdata = AppendTo[ eldhatdata, {time,eldhat*1000}];
     rhatdata = AppendTo[ rhatdata, {time,rhat}];
     rhatddata = AppendTo[ rhatddata, {time,rhatc ';
     azelupdate[ ];

(* third through fourth samples    *)

     Do[{
          time = data[[i,1]];
          rhat = ((i-1)*rhat + data[[i,4]])/i;
          rhatd = (data[[i,4]] - data[[i-1,4]])/dtr;
          azdhat = ((i-1)*azdhat + data[[i,2]]/1000)/i;
          eldhat = ((i-1)*eldhat + data[[i,3]]/1000)/i;
          azdhatdata = AppendTo[ azdhatdata, {time,azdhat*1000}];
          eldhatdata = AppendTo[ eldhatdata, {time,eldhat*1000}];
          rhatdata = AppendTo[ rhatdata, {time,rhat}];
          rhatddata = AppendTo[ rhatddata, {time,rhatd}];
          azelupdate[ ]},
          {i, 3, 4}];

   Do[{
          time = data[[i,1]];
          c1 = 1 - kpr;
          c2 = dtr*c1;
          c3 = 1 - kvr*dtr;
          term1 = 1 - kpe;
          term2 = 1 - kpa;
          rhat = c1*rhat + c2*rhatd + kpr*data[[i,4]];
          term3 = (eldhat^2 + azdhat^2)*dtr;
          rhatd = rhat*(term3 - kvr) + rhatd*c3 + data[[i,4]]*kvr;
          eldhat = term1*(1 - 2*rhatd*dtr/rhat)*eldhat +
                    kpe*data[[i,3]]/1000;
          azdhat = term2*(1 - 2*rhatd*dtr/rhat)*azdhat +
                   kpa*data[[i,2]]/1000;
          azdhatdata = AppendTo[ azdhatdata, {time,azdhat*1000}];
          eldhatdata = AppendTo[ eldhatdata, {time,eldhat*1000}];
          rhatdata = AppendTo[ rhatdata, {time,rhat}];
          rhatddata = AppendTo[ rhatddata, {time,rhatd}];
          azelupdate[ ]},
       {i, 5, Length[data]}];
```

```
aztheo = <<aztheo.dat;
eltheo = <<eltheo.dat;
rngtheo = <<rngtheo.dat;
rngdtheo = <<rngdtheo.dat;
azcrtheo = <<azcrth.dat;
elcrtheo = <<elcrth.dat;
azcorrdata = Rest[azcorrdata];
elcorrdata = Rest[elcorrdata];
azcrdiff = Table[{azcorrdata[[i,1]],azcorrdata[[i,2]]-
           azcrtheo[[i,2]]}, {i,1,Length[azcorrdata]}];
elcrdiff = Table[{elcorrdata[[i,1]],elcorrdata[[i,2]]-
           elcrtheo[[i,2]]}, {i,1,Length[elcorrdata]}];
plotterca[azcorrdata,azcrtheo,
         "azcorr & theo. azcorr for scenario:",
         "t(sec)",
         "mrad",
         {50,100}];
plotterda = ListPlot[azcrdiff, PlotJoined -> True,
         PlotStyle -> {Thickness[0.002],PointSize[0.002]},
         PlotRange -> {-5,5},
         AxesLabel -> {"t(sec)","mrad"},
         PlotLabel -> "difference between azcorr and theo.
                       azcorr"],
plotterce[elcorrdata,elcrtheo,
         "elcorr & theo elcorr for scenario:",
         "t(sec)",
         "mrad/s",
         {10,20}];
plotterde = ListPlot[elcrdiff, PlotJoined -> True,
         PlotStyle -> {Thickness[0.002], PointSize[0.002]},
         PlotRange -> {-5,5},
         AxesLabel -> {"t(sec)","mrad"},
         PlotLabel -> "difference between elcorr and theo.
                       elcorr"],
plottera[azdhatdata,aztheo,
         "azdhat & theo. az. rate for scenario:",
         "t(sec)",
         "mrad/s",
         {-80,-30}];
plottere[eldhatdata,eltheo,
         "eldhat & theo el. rate for scenario:",
         "t(sec)",
         "mrad/s",
         {-12,4}];
plotterr[rhatdata,rngtheo,
         "rdhat & theo range for scenario:",
         "t(sec)",
         "feet",
         {3000,5000}];
plotterd[rhatddata,rngdtheo,
         "rhatd & theo range rate for scenario:",
         "t(sec)",
```

```
                "ft/s",
                {-200,200}];
}]
```

COMPUTER PROGRAM: KALMAN.BAS

PURPOSE: Filters noisy azimuth/elevation rate and range data using adaptive Kalman filtering.

RESULTS: Filtered azimuth/elevation rates, range/range rate and azimuth/elevation corrections

```
10 OPEN "I", #1, "B:NOISE1.GUN"
20 OPEN "O", #2, "B:AZDOT1.GUN"
30 OPEN "O", #3, "B:ELDOT1.GUN"
40 OPEN "O", #4, "B:AZCORR1N.GUN"
50 OPEN "O", #5, "B:ELCORR1N.GUN"
60 P11 = 10000
70 P22 = 300
80 PA = .02
90 PE = .003
100 RR = 10
110 RA = .006
120 RE = .01
130 QR = .7
140 QA = .006
150 QE = .0001
160 TIME = 0
170 DTR = .22
180 DTAE = .0275
190 I = 1
200 RHAT = SENRNG
210 RHATD = 0
220 AZDHAT = 0
230 ELDHAT = 0
240 INPUT #1, TIME, GXHAT, GYHAT, SENRNG
250 PBA = (1-2*DTAE*RHATD/RHAT)^2*PA+DTAE^2*QA
260 PBE = (1-2*DTAE*RHATD/RHAT)^2*PE+DTAE^2*QE
270 KPA = PBA/(PBA+RA)
280 KPE = PBE/(PBE+RE)
290 PA = (1-KPA)*PBA
300 PE = (1-KPE)*PBE
310 TERM1 = 1-KPE
320 TERM2 = 1-KPA
330 ELDHAT = TERM1*(1-2*RHATD*DTAE/RHAT)*ELDHAT+KPE*GYHAT/1000
340 AZDHAT = TERM2*(1-2*RHATD*DTAE/RHAT)*AZDHAT+KPA*GXHAT/1000
350 IF ((I-1) MOD 8) = 0, THEN GOSUB 560
360 ELCORR = RHAT*ELDHAT/2840
370 ELCORR = -ATN(ELCORR/SQR(1-ELCORR^2))
380 AZCORR = RHAT*AZDHAT/2840
390 AZCORR = -ATN(AZCORR/SQR(1-AZCORR^2))
400 VP = RHATD+2840*COS(ELCORR)*SIN(AZCORR)
410 WCDA = .000005*RHAT^2-.036938*RHAT+700.429
420 BETA = 3E-09*RHAT^2-4.706E-05*RHAT+1.0216
```

```
430 RHO = .076479
440 C = 1.5*RHO/WCDA
450 TOF = RHAT/VP/(1-C*RHAT/4)^2
460 ELGD = 16.085*(BETA*TOF)^2/RHAT
470 ELCORR = ELCORR+ELGD
480 PRINT TIME, AZDHAT*1000, ELDHAT*1000, AZCORR*1000, ELCORR*1000
490 PRINT #2, TIME, AZDHAT*1000
500 PRINT #3, TIME, ELDHAT*1000
510 PRINT #4, TIME, AZCORR*1000
520 PRINT #5, TIME, ELCORR*1000
530 I = I + 1
540 GOTO 240
550 END
560 PB11 = P11+DTR*(P12+P21+P22*DTR)
570 PB12 = P12+DTR*(P22+(P11+DTR*P21)*OMEGA2)
580 PB21 = P21+DTR*(P22+(P11+DTR*P12)*OMEGA2)
590 PB22 = P22+QR*DTR^2+OMEGA2*DTR*(P21+P12+DTR*OMEGA2*P11)
600 DENOM = PB11+RR
610 KPR = PB11/DENOM
620 KVR = PB21/DENOM
630 P11 = PB11*RR/DENOM
640 P12 = PB12*RR/DENOM
650 P21 = PB21*RR/DENOM
660 P22 = PB22-PB21*PB12/DENOM
670 C1 = 1-KPR
680 C2 = DTR*C1
690 C3 = 1-KVR*DTR
700 TERM3 = OMEGA2*DTR
710 RHAT = C1*RHAT+C2*RHATD+KPR*SENRNG
720 RHATD = RHAT*(TERM3-KVR)+RHATD*C3+KVR*SENRNG
730 RETURN
```

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

Universal Energy Systems, Inc.

FINAL REPORT FOR THE PERIOD 1/1/91 TO 12/31/91

ROBUST EIGENSTRUCTURE ASSIGNMENT WITH APPLICATION TO MISSILE CONTROL

Prepared by:	Kenneth M. Sobel, Ph.D.
Academic Rank:	Associate Professor
Department and	Electrical Engineering
University:	The City College of New York
Research Location:	WL/MNAG Eglin AFB, FL 32542-5434
USAF Researcher:	Dr. J. Cloutier
Date:	31 Dec 91
Contract No:	F49620-88-C-0053

ABSTRACT

A new sufficient condition for the robust stability of a linear time invariant continuous time system subject to linear time varying structured state space uncertainty is presented. A robust eigenstructure assignment design is computed for the Extended Medium Range Air to Air Missile. The new design method uses the MATLAB™ Optimization Toolbox to minimize the integral of the roll rate with constraints on the real part of the dutch roll and roll mode eigenvalues, the damping ratios of the dutch roll and roll mode, the aileron and rudder deflection rates, and the sufficient condition for robust stability.

Next, we extend eigenstructure assignment to linear time invariant plants which are represented by the so-called unified delta model which is valid both for continuous time and sampled data operation of the plant. We propose a sufficient condition for the robust stability of a linear time invariant unified delta plant subject to linear time invariant structured state space uncertainty. A robust sampled data design is computed for the Extended Medium Range Air to Air Missile by minimizing the integral of the roll rate with constraints on selected eigenvalues, actuator deflection rates, and the sufficient condition for robust stability.

Finally, we extend the robust stability result proposed by Sobel et. al. (1989) and the new continuous time result described in this report to continuous time systems with a time delay.

LIST OF FIGURES

Fig. 1. Sideslip angle (continuous time; $\beta(0)=1$ deg)

Fig. 2. Yaw rate (continuous time; $\beta(0)=1$ deg)

Fig. 3. Roll rate (continuous time; $\beta(0)=1$ deg)

Fig. 4. Integrated roll rate (continuous time; $\beta(0)=1$ deg)

Fig. 5. Rudder deflection (continuous time; $\beta(0)=1$ deg)

Fig. 6. Aileron deflection (continuous time; $\beta(0)=1$ deg)

Fig. 7. Roll rate step response (continuous time)

Fig. 8. Rudder deflection for step response (continuous time)

Fig. 9. Aileron deflection for step response (continuous time)

Fig. 10. Sideslip angle (delta model; $\beta(0)=1$ deg)

Fig. 11. Yaw rate (delta model; $\beta(0)=1$ deg)

Fig. 12. Roll rate (delta model; $\beta(0)=1$ deg)

Fig. 13. Integrated roll rate (delta model; $\beta(0)=1$ deg)

Fig. 14. Rudder deflection (delta model; $\beta(0)=1$ deg)

Fig. 15. Aileron deflection (delta model; $\beta(0)=1$ deg)

Fig. 16. Sideslip angle (delta model; β_g = 1-cosine)

Fig. 17. Yaw rate (delta model; β_g = 1-cosine)

Fig. 18. Roll rate (delta model; β_g = 1-cosine)

Fig. 19. Integrated roll rate (delta model; β_g = 1-cosine)

Fig. 20. Rudder deflection (delta model; β_g = 1-cosine)

Fig. 21. Aileron deflection (delta model; β_g = 1-cosine)

1. INTRODUCTION

Recently, Sobel and Cloutier (1991) applied eigenstructure assignment to the design of an autopilot for the Extended Medium Range Air to Air Technology (EMRAAT) missile. An important difference between this application and other eigenstructure assignment applications which have appeared in the literature is that the lateral dynamics of the EMRAAT missile does not have a well defined dutch roll mode. Therefore, eigenstructure assignment is utilized not only for mode decoupling, but also to create distinctly separate dutch roll and roll modes. Sobel and Cloutier (1991) use the approach suggested by Andry et. al. (1983) in which the i-th desired eigenvector v_i^d is chosen for mode decoupling. Then, the i-th achievable eigenvector v_i^a is chosen as the projection of v_i^d onto the so-called achievability subspace. Sobel and Cloutier (1991) show that their design achieves improved decoupling between an initial sideslip angle and the integrated roll rate (which is approximately equal to the bank angle) when compared to a linear quadratic regulator design proposed by Bossi and Langehough (1988). However, the design of Sobel and Cloutier (1991) does not consider that the missile's aerodynamic parameters are uncertain.

A sufficient condition for the robust stability of a linear time invariant system subject to linear time varying structured state space uncertainty has been proposed by Sobel et. al. (1989). This result, which is based upon the Gronwall lemma, ensures robust stability if the nominal eigenvalues lie to the left of a vertical line in the complex plane. This line is determined by the maximum eigenvalue of a matrix

which involves the product of the uncertainty structure, the nominal closed loop modal matrix, and the inverse of the nominal closed loop modal matrix.

In this report, we present a new sufficient condition for the robust stability of a linear time invariant system subject to linear time varying structured state space uncertainty. This new robustness condition is a sum of terms each of which involves the i-th right eigenvector, the i-th left eigenvector, and the real part of the i-th eigenvalue. We also show that this new robustness condition is less conservative than the earlier result of Sobel et. al. (1989).

We design a robust controller for the lateral dynamics of the EMRAAT missile and this new design is compared to the earlier orthogonal projection eigenstructure assignment design. The earlier design, which was proposed by Sobel and Cloutier (1991), does not satisfy either sufficient condition for robust stability. The new design method proposed in this paper uses the MATLAB™ Optimization Toolbox (Grace 1990) to minimize the integrated roll rate with constraints on the real part of the dutch roll and roll modes, the damping ratios of the dutch roll and roll modes, the aileron and rudder deflection rates, and the new sufficient condition for robust stability. This design satisfies the new robustness condition while also yielding an improved transient response as compared to the design of Sobel and Cloutier (1991).

We extend eigenstructure assignment to linear time invariant plants which are represented by the so-called unified delta model which is valid both for continuous time and sampled data operation of the plant.

We show that the eigenvectors of the delta model are identical to the eigenvectors of the continuous time plant and an expression is derived for the eigenstructure assignment feedback gain matrix for the delta model. We show that in the limit as the sampling period delta goes to zero, the delta feedback gain approaches the continuous time feedback gain. We propose a sufficient condition for the robust stability of a linear time invariant unified delta plant subject to linear time invariant structured state space uncertainty.This yields a new unified robustness condition which is applicable to both continuous time and sampled data operation. A robust sampled data design is computed for the Extended Medium Range Air to Air Missile by minimizing the integral of the roll rate with constraints on selected eigenvalues, actuator deflection rates, and the sufficient condition for robust stability. The robust design is compared with an orthogonal projection eigenstructure assignment design.

Finally, we extend the robust stability results proposed by Sobel et. al. (1989) and the new results described in this report to continuous time systems with a time delay. We proof new sufficient conditions for the robust stability of a linear time invariant continuous time system with a time delay which is subject to linear time varying structured state space uncertainty.

2. ROBUST EIGENSTRUCTURE ASSIGNMENT FOR CONTINUOUS TIME SYSTEMS

2.1 Problem Formulation

Consider a nominal linear time-invariant multi-input multi-output system described by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1a}$$

$$y(t) = Cx(t) \tag{1b}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the input vector, $y \in \mathbb{R}^r$ is the output vector, and A, B, C, are constant matrices.

Suppose that the nominal system is subject to linear time-varying uncertainties in the entries of A, B described by $\Delta A(t)$ and $\Delta B(t)$, respectively. We shall assume that the entries of $\Delta A(t)$ and $\Delta B(t)$ are uniformly continuous for $t \in (-\infty,\infty)$. Then, the system with uncertainty is given by

$$\dot{x}(t) = Ax(t) + Bu(t) + \Delta A(t)x(t) + \Delta B(t)u(t) \tag{2a}$$

$$y(t) = Cx(t) \tag{2b}$$

Further, suppose that bounds are available on the absolute values of the maximum variations in the elements of $\Delta A(t)$ and $\Delta B(t)$. That is,

$$|\Delta a_{ij}(t)| \le (a_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,n \tag{3a}$$

$$|\Delta b_{ij}(t)| \le (b_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,m \tag{3b}$$

Define $\Delta A^+(t)$ and $\Delta B^+(t)$ as the matrices obtained by replacing the entries of $\Delta A(t)$ and $\Delta B(t)$ by their absolute values. Also, define A_{max} and B_{max} as the matrices with entries $(a_{ij})_{max}$ and $(b_{ij})_{max}$, respectively. Then,

$$\{\Delta A(t):\ \Delta A^{+}(t) \le A_{max}\} \tag{4a}$$

and

$$\{\Delta B(t):\ \Delta B^{+}(t) \le B_{max}\} \tag{4b}$$

where "≤" is applied element by element to matrices and $A_{max} \in \mathbb{R}_{+}^{n\times n}$, $B_{max} \in \mathbb{R}_{+}^{n\times m}$ where $\mathbb{R}_{+}$ is the set of non-negative numbers.

Consider the constant gain output feedback control law described by

$$u(t) = Fy(t) \tag{5}$$

Then, the nominal closed loop system is given by

$$\dot{x}(t) = (A + BFC)x(t) \tag{6}$$

and the uncertain closed loop system is given by

$$\dot{x}(t) = (A + BFC)x(t) + [\Delta A(t) + \Delta B(t)FC]x(t) \tag{7}$$

Finally, the stability robustness problem can be stated as follows: Given a feedback gain matrix $F \in \mathbb{R}^{m\times r}$ such that the nominal closed loop system exhibits desirable dynamic performance, determine if the uncertain closed loop system is asymptotically stable for all $\Delta A(t)$ and $\Delta B(t)$ described by Eq.(4).

2.1 Robustness Results

The solution proposed by Sobel et. al. (1989) for the stability robustness problem is described by the following theorem.

Theorem 1: Suppose that F is such that the nominal closed loop system described by Eq.(6) is asymptotically stable with a non-defective modal matrix. Then, the uncertain closed loop system given by Eq.(7) is asymptotically stable for all $\Delta A(t)$ and $\Delta B(t)$ described by Eq.(4) if

$$\alpha > \lambda_{max} \, [(M^{-1})^{+}[A_{max} + B_{max}(FC)^{+}]M^{+}] \tag{8}$$

where

$$\alpha = -\max_{i} \, Re[\lambda_i(A+BFC)]$$

and where M is a modal matrix of (A+BFC) and $\lambda_{max}(\cdot)$ of a non-negative matrix denotes the real non-negative eigenvalue $\lambda_{max} \geq 0$ such that $\lambda_{max} \geq |\lambda_i|$ for all eigenvalues λ_i.

The above theorem describes a sufficient condition for the robust stability in terms of the eigenstructure of the nominal closed loop system. Robust stability is ensured provided that the nominal closed loop eigenvalues lie to the left of a vertical line in the complex plane which is determined by a norm involving the structure of the uncertainty and the nominal closed loop modal matrix.

We remark that the sufficient condition given by Eq.(8) can be rewritten as shown below.

$$\left(\frac{1}{\alpha}\right) \cdot \lambda_{max} \, [(M^{-1})^{+}[A_{max} + B_{max}(FC)^{+}]M^{+}] < 1 \tag{9}$$

This alternate form of the stability robustness condition will be useful later when we compare it with the new robustness condition which is derived in this report.

We now present a new sufficient condition for robust stability of a linear time invariant system subject to linear time varying structured state space uncertainty which is described by the following theorem.

Theorem 2: Suppose that F is such that the nominal closed loop system described by Eq.(6) is asymptotically stable with a non-defective modal matrix. Then, the uncertain closed loop system given by Eq.(7) is

asymptotically stable for all $\Delta A(t)$ and $\Delta B(t)$ described by Eq.(4) if

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} \left[A_{max} + B_{max}(FC)^+\right] \right\} < 1 \tag{10}$$

where

$$\alpha_i = -\mathrm{Re}[\lambda_i(A+BFC)]$$

and where λ_i is the i-th eigenvalue of (A+BFC) with v_i and w_i the corresponding right and left eigenvectors, respectively; and where $(\cdot)^*$ denotes the complex conjugate transpose.

Proof: The uncertain closed loop plant may be written as

$$\dot{x}(t) = A_c x(t) + \Delta A_c(t) x(t) \tag{11}$$

where

$$A_c = A + BFC$$

and

$$\Delta A_c(t) = \Delta A(t) + \Delta B(t) FC$$

which has a solution given by

$$x(t) = \exp(A_c t)x(0) + \int_0^t \exp[A_c(t-\tau)]\Delta A_c(\tau)x(\tau)d\tau \tag{12}$$

Next, use the real, positive, diagonal transformation

$$x(t) = D^{-1}z(t) \tag{13}$$

and the property that

$$\exp(DA_c D^{-1}t) = D\exp(A_c t)D^{-1} \tag{14}$$

and

$$\exp(A_c t) = M\exp(\Lambda t)M^{-1} = \sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} \tag{15}$$

to obtain

$$z(t) = D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0) + \int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-\tau)} \Delta A_c(\tau) D^{-1} z(\tau) d\tau \tag{16}$$

where M is a modal matrix of A_c; λ_i is the i-th eigenvalue of A_c with v_i and w_i the corresponding right and left eigenvectors, respectively; Λ is a diagonal matrix with the λ_i on the diagonal; and $(\cdot)^*$ denotes complex conjugate transpose.

Note that

$$\|z(t)\| \to 0 \text{ implies that } \|x(t)\| \to 0 \tag{17}$$

Next, apply the absolute value operator, denoted by $(\cdot)^+$, to both sides of Eq.(16) where "+" and "≤ " are applied element by element to vectors and matrices.

$$z^+(t) \le \left[D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0)\right]^+ + \left[\int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-\tau)} \Delta A_c(\tau) D^{-1} z(\tau) d\tau\right]^+ \tag{18}$$

$$\le D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i t} D^{-1} z^+(0) + \int_0^t \left[D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-\tau)} \Delta A_c(\tau) D^{-1} z(\tau)\right]^+ d\tau \tag{19}$$

$$\le D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i t} D^{-1} z^+(0) + \int_0^t D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-\tau)} A_{cmax} D^{-1} z^+(\tau) d\tau \tag{20}$$

where $\alpha_i = -\mathrm{Re}(\lambda_i)$, $A_{cmax} = A_{max} + B_{max}(FC)^+$ and where we have used the property that $[\exp(\lambda_i t)]^+ = \exp(-\alpha_i t)$.

Next, integrate both sides of Eq.(20) to obtain

$$\int_0^\infty z^+(t)dt \le D\sum_{i=1}^{n} (v_i w_i^*)^+ \int_0^\infty e^{-\alpha_i t} dt \; D^{-1} z^+(0)$$

$$+ \int_{t=0}^{\infty} \int_{\tau=0}^{t} D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-\tau)} A_{cmax} D^{-1} z^+(\tau) d\tau dt \tag{21}$$

Consider the double integral in Eq.(21) which may be written as

$$\lim_{R\to\infty}\left[\int_{t=0}^{R}\int_{\tau=0}^{t} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i(t-\tau)} A_{cmax}D^{-1}z^+(\tau)d\tau dt\right];\ 0\le\tau<t \qquad (22)$$

The order of integration in Eq.(22) may be interchanged because all of the functions inside the integrals are continuous functions of t and τ (Churchill 1972). Thus, Eq.(22) is equal to

$$\lim_{R\to\infty}\left[\int_{\tau=0}^{t}\int_{t=0}^{R} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i(t-\tau)} A_{cmax}D^{-1}z^+(\tau)dt d\tau\right];0\le\tau<t \qquad (23)$$

Now use the change of variables given by $\gamma=t-\tau$, $d\gamma=dt$, $\gamma>0$. Then, Eq.(23) is equal to

$$\lim_{R\to\infty}\left[\int_{\tau=0}^{t}\int_{\gamma=-\tau}^{R-\tau} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax}D^{-1}z^+(\tau)d\gamma d\tau\right];0\le\tau<t,\ \gamma>0 \qquad (24)$$

$$\le \lim_{R\to\infty}\left[\int_{\tau=0}^{t}\int_{\gamma=0}^{R} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax}D^{-1}z^+(\tau)d\gamma d\tau\right];\alpha_i>0 \qquad (25)$$

$$= \lim_{R\to\infty}\left[\int_{\gamma=0}^{R} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax}D^{-1}d\gamma\int_{\tau=0}^{t} z^+(\tau)d\tau\right];\alpha_i>0 \qquad (26)$$

$$\le \lim_{R\to\infty}\left[D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{-\alpha_i}(e^{-\alpha_i R}-1)\,A_{cmax}D^{-1}\int_0^{\infty} z^+(\tau)d\tau\right];\alpha_i>0 \qquad (27)$$

Evaluating the limit of the term outside the integral in Eq.(27), note that τ is now a dummy variable of integration, recall that the α_i's are positive, and substitute the result into Eq.(21) to obtain

$$\int_0^{\infty} z^+(t)dt \le D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}D^{-1}z^+(0) + D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax}D^{-1}\int_0^{\infty} z^+(t)dt \qquad (28)$$

Take norms in Eq.(28) and rearrange to obtain

$$\left\| \int_0^\infty z^+(t)dt \right\| \le \frac{\left\| D\sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} D^{-1} z^+(0) \right\|}{1 - \left\| D\sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} A_{cmax} D^{-1} \right\|} \tag{29}$$

Thus, $\left\| \int_0^\infty z^+(t)\, dt \right\| < \infty$ which implies that $\int_0^\infty \|z^+(t)\|\, dt < \infty$ if

$$\left\| D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} [A_{max} + B_{max} (FC)^+] D^{-1} \right\| < 1 \tag{30}$$

Note that x(t) and z(t) are uniformly continuous on $(0,\infty)$ because of the linearity of the uncertain closed loop plant which together with Eq.(30) implies that $\|z^+(t)\| \to 0$ as $t\to\infty$ (Hsu and Meyer 1968). This implies that $\|x(t)\| \to 0$ as $t\to \infty$ which proves that the linear uncertain closed loop plant is asymptotically stable.

Finally, Perron weightings may be used for the matrix D in Eq.(30) in the same manner as shown by Sobel et. al. (1989) for an earlier robustness result. Thus, to reduce conservatism, Eq.(29) may be replaced by

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} [A_{max} + B_{max}(FC)^+] \right\} < 1 \tag{31}$$

where $\lambda_{max}(\cdot)$ of a non-negative matrix denotes the real non-negative eigenvalue $\lambda_{max} \ge 0$ such that $\lambda_{max} \ge |\lambda_i|$ for all eigenvalues λ_i.

Q.E.D.

We remark that the quantity $|w_i^* v_i|$ is defined by Golub and Van Loan (1983) to be the condition number of the i-th eigenvalue. This condition number is a measure of the sensitivity of the i-th eigenvalue to incremental perturbations in the matrix A+BFC. The quantity $v_i w_i^*$, which appears in Eq.(10), is weakly related to this eigenvalue condition number. Thus, the new stability robustness condition of Eq.(10) seems to have the heuristic interpretation that robustness can be achieved by having the sensitive eigenvalues far into the left half complex plane and by having the eigenvalues which are near the imaginary axis to be insensitive. Such an interpretation was not possible for the result of Sobel et. al. (1989) because their stability robustness condition, which is given by Eq.(9), is in terms of the spectrum as a whole rather than the individual eigenvalues.

The new stability robustness condition of Eq.(10) is less conservative than the earlier condition of Sobel et. al. (1989). This property is described by the following theorem.

Theorem 3: The left hand side of Eq.(10) is less than or equal to the left hand side of Eq.(9). Mathematically,

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i}[A_{max}+B_{max}(FC)^+]\right\} \le \left(\frac{1}{\alpha}\right)\lambda_{max}\left\{(M^{-1})^+[A_{max}+B_{max}(FC)^+]M^+\right\} \tag{32}$$

Proof: We element by element bound the matrix on the left hand side of Eq.(32) and recall the definition that $A_{cmax} = A_{max} + B_{max}(FC)^+$ to obtain

$$\sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} A_{cmax} \le \sum_{i=1}^{n} \frac{v_i^+ (w_i^*)^+}{\alpha_i} A_{cmax} \tag{33}$$

$$\le \sum_{i=1}^{n} \frac{v_i^+ (w_i^*)^+}{\alpha} A_{cmax} \tag{34}$$

$$\le \left(\frac{1}{\alpha}\right) \cdot \left[v_1^+, v_2^+, \ldots, v_n^+ \right] \begin{bmatrix} (w_1^*)^+ \\ (w_2^*)^+ \\ \vdots \\ (w_n^*)^+ \end{bmatrix} \cdot A_{cmax} \tag{35}$$

$$= \left(\frac{1}{\alpha}\right) \cdot M^+ (M^{-1})^+ A_{cmax} \tag{36}$$

Now, use the result that if $B^+ \le A^+$, then

$$\lambda_{max}(B^+) = \|B^+\|_{2D_{B^+}} \le \|B^+\|_{2D_{A^+}} \le \|A^+\|_{2D_{A^+}} = \lambda_{max}(A^+) \tag{37}$$

where $\|\cdot\|_{2_{D_{B^+}}}$ and $\|\cdot\|_{2_{D_{A^+}}}$ are D-weighted two norms with Perron weights for B^+ and A^+, respectively.

Then, applying Eq.(37) to Eq.(36) yields

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} A_{cmax} \right\} \le \lambda_{max} \left\{ \left(\frac{1}{\alpha}\right) \cdot M^+ (M^{-1})^+ A_{cmax} \right\} \tag{38}$$

Finally, use the result (Ogata 1987) that if A and B are square matrices, then

$$\lambda_i(AB) = \lambda_i(BA) \tag{39}$$

to obtain

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{\alpha_i} A_{cmax} \right\} \le \lambda_{max} \left\{ \left(\frac{1}{\alpha}\right) \cdot (M^{-1})^+ A_{cmax} M^+ \right\} \qquad (40)$$

Q.E.D.

2.1 Application to the EMRAAT Missile

Consider the Extended Medium Range Air to Air Technology (EMRAAT) bank-to-turn missile which is described by Bossi and Langehough (1988). A sixth order model of the yaw/roll dynamics at a 10 degree angle of attack is considered with state vector, control vector, and measurement vector given by $x = [\beta, r, p, p_I, \delta_r, \delta_a]^T$, $u = [\delta_r, \delta_a]^T$, and $y = [\beta, r, p, p_I]^T$, respectively. Here β is the sideslip angle (deg), r is yaw rate (deg/sec), p is roll rate (deg/sec), p_I is integrated roll rate (deg), δ_r is rudder deflection (deg), and δ_a is aileron deflection (deg). The state space matrices A, B, and C are shown below:

$$A = \begin{bmatrix} -.5007 & -.9945 & .1736 & 0 & .109 & .00691 \\ 16.83 & -.5748 & .01233 & 0 & -132.8 & 27.19 \\ -322.7 & .3208 & -2.099 & 0 & -1620.0 & -1240.0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -179 & 0 \\ 0 & 0 & 0 & 0 & 0 & -179 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 179 & 0 \\ 0 & 0 & 0 & 0 & 0 & 179 \end{bmatrix}^T$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Sobel and Cloutier (1991) use eigenstructure assignment with an orthogonal projection. The desired dutch roll and roll mode eigenvalues are achieved exactly because four measurements are available for feedback. The desired dutch roll eigenvectors are chosen to yield a complex mode which is composed of sideslip angle and yaw rate with no coupling to roll rate and integrated roll rate. The desired roll mode eigenvectors are chosen to yield a complex mode which is composed of roll rate and integrated roll rate with no coupling to sideslip angle and yaw rate. Then, the achievable eigenvectors are computed by using the orthogonal projection of the i-th desired eigenvector v_i^d onto the subspace which is spanned by the columns of $(\lambda_i I-A)^{-1}B$. The closed loop eigenvalues and the feedback gain matrix are shown in Table 1.

TABLE 1. COMPARISON OF EMRAAT DESIGNS

	Closed Loop Eigenvalues*	Feedback Gain Matrix
Eigen-structure Assignment Design of Sobel and Cloutier (1991)	$\lambda_{dr}=-23.98\pm j17.99$ $\lambda_{roll}=-10.01\pm j9.98$ $\lambda_{act} = -132.1$ $\lambda_{act} = -161.1$	$\begin{array}{c} \begin{array}{cccc} \beta & r & p & p_I \end{array} \\ \begin{bmatrix} -4.19 & .233 & .00374 & .731 \\ 2.89 & -.290 & .00631 & -.812 \end{bmatrix} \end{array} \begin{array}{c} \delta_{rc} \\ \delta_{ac} \end{array}$
Robust Design	$\lambda_{dr}=-14.94\pm j16.68$ $\lambda_{roll}=-40.55\pm j92.90$ $\lambda_{act} = -150.3$ $\lambda_{act} = -99.3$	$\begin{array}{c} \begin{array}{cccc} \beta & r & p & p_I \end{array} \\ \begin{bmatrix} -2.82 & .162 & .0127 & 1.18 \\ .770 & -.180 & .062 & 3.14 \end{bmatrix} \end{array} \begin{array}{c} \delta_{rc} \\ \delta_{ac} \end{array}$

*Eigenvalues are computed by using feedback gains which are rounded to three significant digits.

In this report, we propose a new robust design which minimizes the integrated roll rate (which is approximately the bank angle) subject to constraints on the real part of the dutch roll and roll modes, the damping ratios of the dutch roll and roll modes, the aileron and rudder deflection rates, and the new sufficient condition for

robust stability. Mathematically, the objective function to be minimized is given by

$$J = \sum_{k=1}^{30} [p_I(kT_1)]^2 \tag{41}$$

where T_1=0.01 sec.

The values of T_1 and k are chosen to include the time interval [0,0.3] during which most of the transient response occurs. Of course, computation of Eq.(41) requires that a linear simulation be performed during each function evaluation of the optimization. The constraints are shown below where ζ is the damping ratio.

$$\mathrm{Re}\ \lambda_{dr} \in [-50,-6] \tag{42}$$

$$\mathrm{Re}\ \lambda_{roll} \in [-50,-6] \tag{43}$$

$$\mathrm{Re}\ \lambda_{rudder} < -100 \tag{44}$$

$$\mathrm{Re}\ \lambda_{aileron} < -100 \tag{45}$$

$$\zeta_{dr} \in [0.4,0.8] \tag{46}$$

$$\zeta_{roll} \in [0.4,0.8] \tag{47}$$

$$|\dot{\delta}_a| < 275 \text{ deg/sec} \tag{48}$$

$$|\dot{\delta}_r| < 275 \text{ deg/sec} \tag{49}$$

$$\lambda_{max}\left\{ \sum_{i=1}^{6} \frac{(v_i w_i^*)^+}{\alpha_i} [A_{max} + B_{max}(FC)^+] \right\} < 0.999 \tag{50}$$

where for illustrative purposes we have chosen $A_{max} = 0.1 \cdot A^+$ and $B_{max} = 0$. The actuator deflection rates are computed from the slopes of the time responses of the deflections during the time interval [0,0.03]. This interval is chosen because the slopes of the deflections are largest during this time interval. Mathematically,

$$|\dot{\delta}_r| = \frac{\max|\delta_r(mT_2)|}{mT_2} \tag{51}$$

$$|\dot{\delta}_a| = \frac{\max|\delta_a(mT_2)|}{mT_2} \tag{52}$$

where T_2=0.001 sec and m=0,1,...,30. The maximum deflection rates chosen for the constraints are well within the expected 400 deg/sec limit for the advanced state of the art electromechanical actuator described by Langehough and Simons (1988).

The parameter vector contains the quantities which may be varied by the optimization. This twelve dimensional vector includes Re λ_{dr}, Im λ_{dr}, Re λ_{roll}, Im λ_{roll}, Re $z_1(1)$, Re $z_1(2)$, Im $z_1(1)$, Im $z_1(2)$, Re $z_3(1)$, Re $z_3(2)$, Im $z_3(1)$, Im $z_3(2)$. Here, the two dimensional complex vectors z_i contain the free eigenvector parameters. That is, the i-th eigenvector v_i may be written as

$$v_i = L_i z_i \tag{53}$$

where the columns of $L_i=(\lambda_i I-A)^{-1}B$ are a basis for the subspace in which the i-th eigenvector must reside. Thus, the free parameters are the vectors z_i rather than the eigenvectors v_i.

The optimization is performed by using subroutine *constr* from the MATLAB™ Optimization Toolbox (Grace 1990) on a 486™ 25MHz personal computer. The optimization is initialized with the design proposed by Sobel and Cloutier (1991) which yields an initial value of 5.0214 for the objective function of Eq.(41), a value of 4.0472 for the left hand side (LHS) of the earlier robustness condition of Eq.(9), and a value of 2.0686 for the LHS of the new robustness condition of Eq.(10). The

optimization is complete after 2698 function evaluations which requires approximately one hour of computation and yields an optimal objective function of 0.0284, a value of 2.4432 for the LHS of the earlier robustness condition of Eq.(9), and a value of 0.999 for the LHS of the new robustness condition of Eq. (10). We observe from Table 1 that the dutch roll mode is dominant in the robust design whereas the roll mode was chosen to be dominant in the design of Sobel and Cloutier (1991). Furthermore, the optimization has assigned the roll mode damping to be the smallest value which is allowed by the constraint of Eq.(47).

The time histories of sideslip angle, yaw rate, roll rate, integrated roll rate, rudder deflection, and aileron deflection to a one degree initial sideslip are shown in figures 1, 2, 3, 4, 5, and 6, respectively. We observe a significant improvement in the integrated roll rate response (which is desired to be zero) when compared to the earlier design of Sobel and Cloutier (1991). The earlier design of Sobel and Cloutier (1991) has a minimum $p_I(t)$ of -.646 deg but the new design of this paper has a minimum $p_I(t)$ of -.112 deg which is an improvement of approximately 80%. We note that this improved response is obtained with both smaller aileron and rudder deflections. Furthermore, it is interesting that the aileron in the design of Sobel and Cloutier (1991) exhibits an initial positive deflection of approximately two degrees before becoming negative, whereas this large positive initial aileron deflection does not appear in the new robust design.

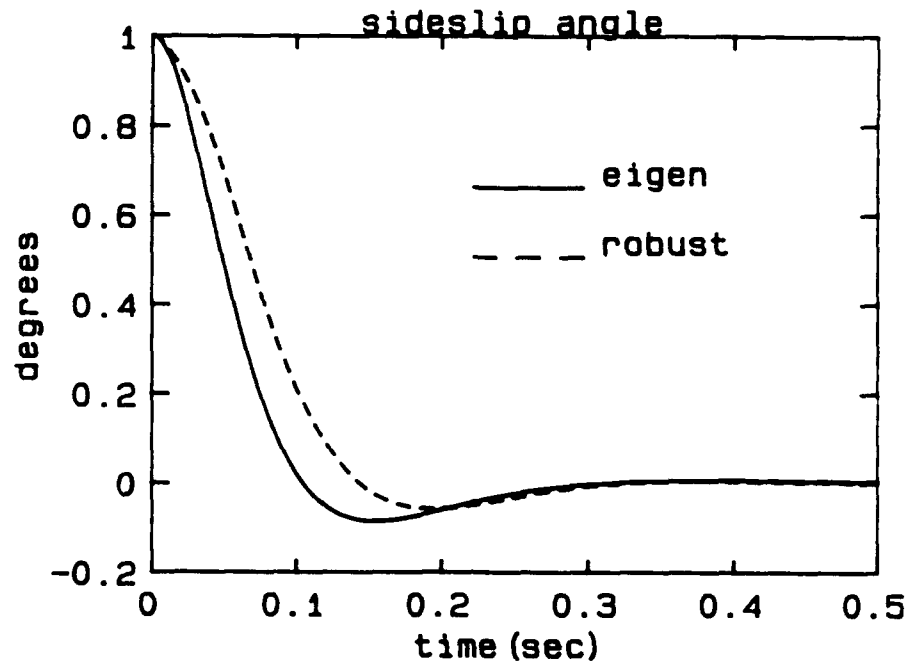


Fig. 1. Sideslip angle (continuous time; $\beta(0)=1$ deg)

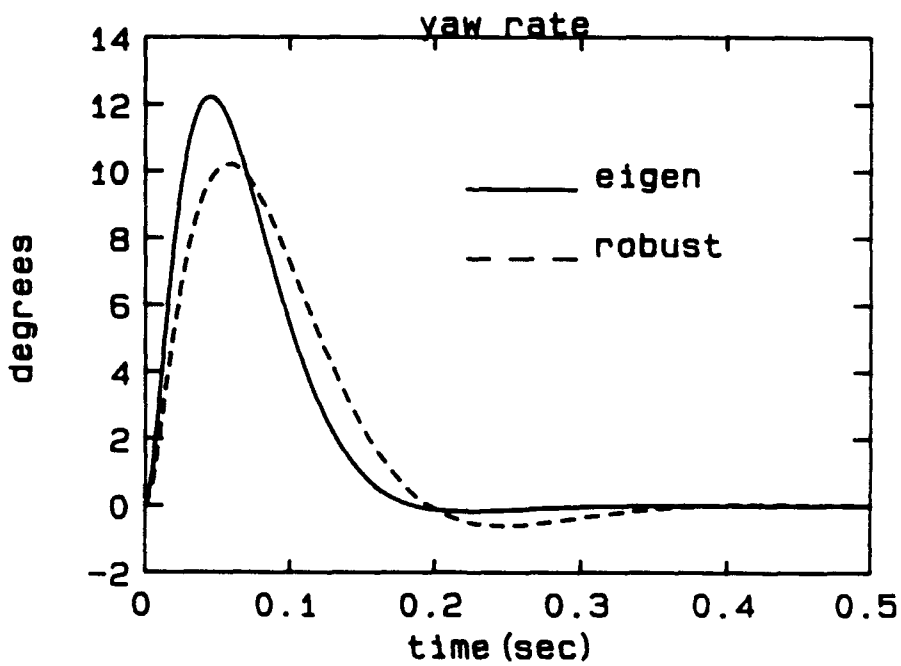


Fig. 2. Yaw rate (continuous time; $\beta(0)=1$ deg)

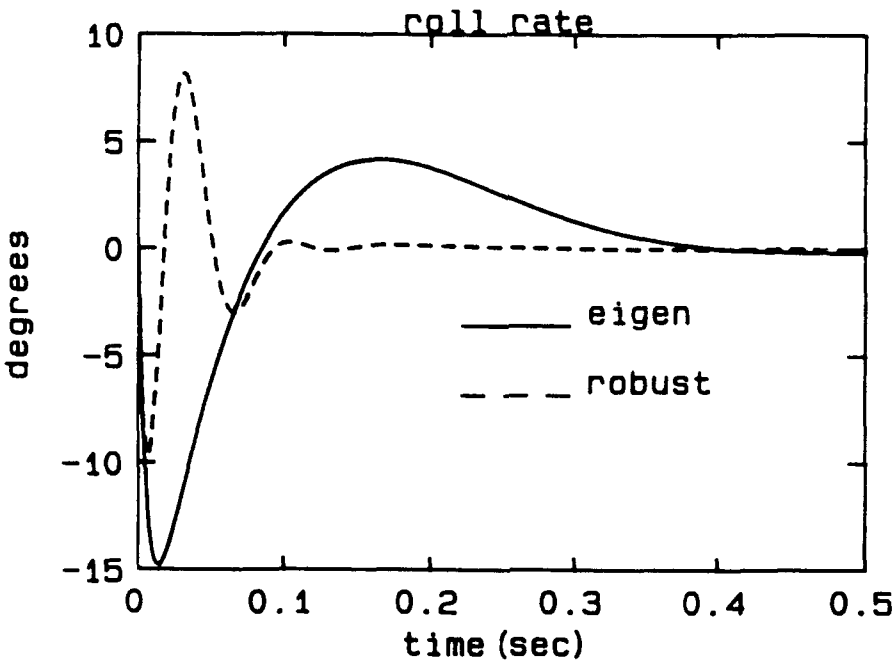


Fig. 3. Roll rate (continuous time; $\beta(0)=1$ deg)

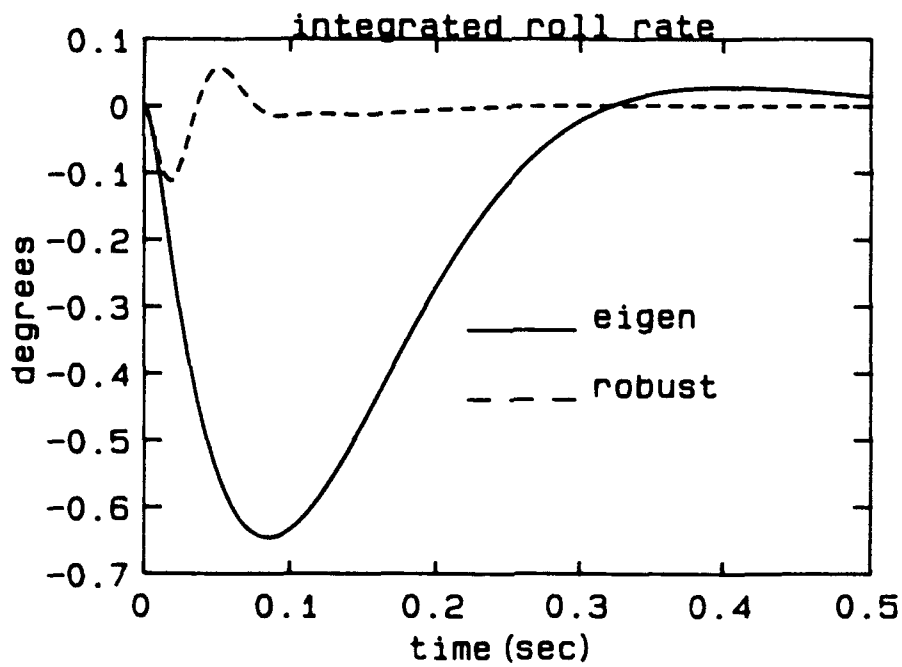


Fig. 4. Integrated roll rate (continuous time; $\beta(0)=1$ deg)

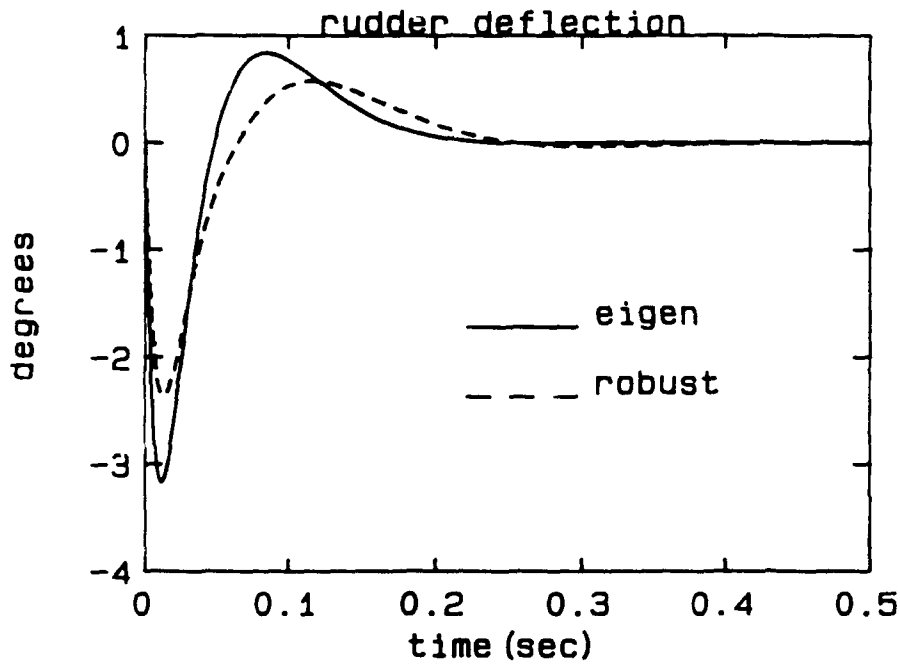


Fig. 5. Rudder deflection (continuous time; $\beta(0)=1$ deg)

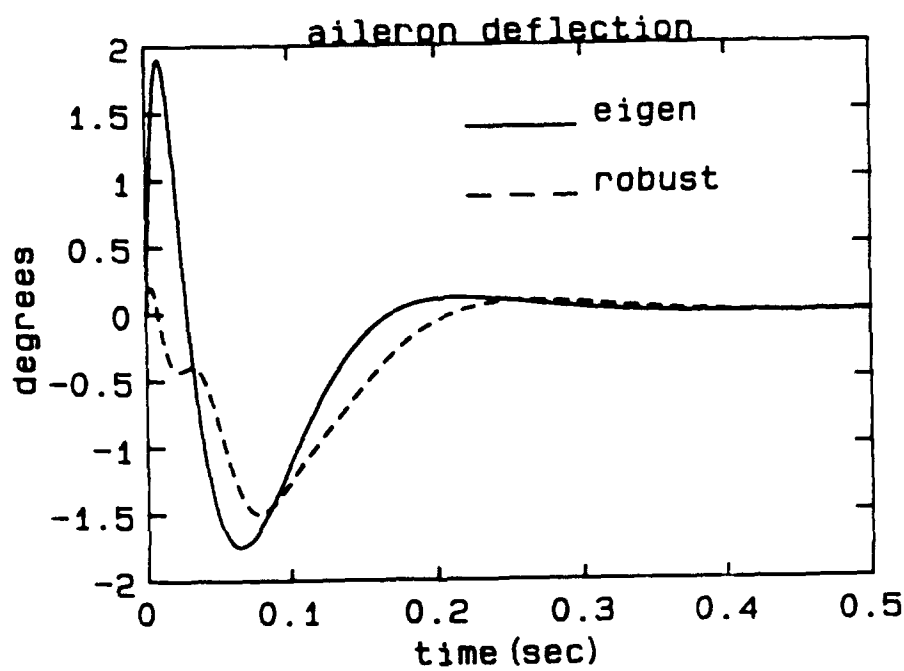


Fig. 6. Aileron deflection (continuous time; $\beta(0)=1$ deg)

Next, the response to a roll rate step of 5 deg/sec is considered. The roll rate, rudder deflection, and aileron deflection are shown in Figures 7, 8, and 9, respectively. The new robust design exhibits a roll rate response with a significantly smaller settling time which is achieved by using larger aileron deflection rates. Nevertheless, these deflection rates of approximately 25 deg/sec are well within the allowable limits.

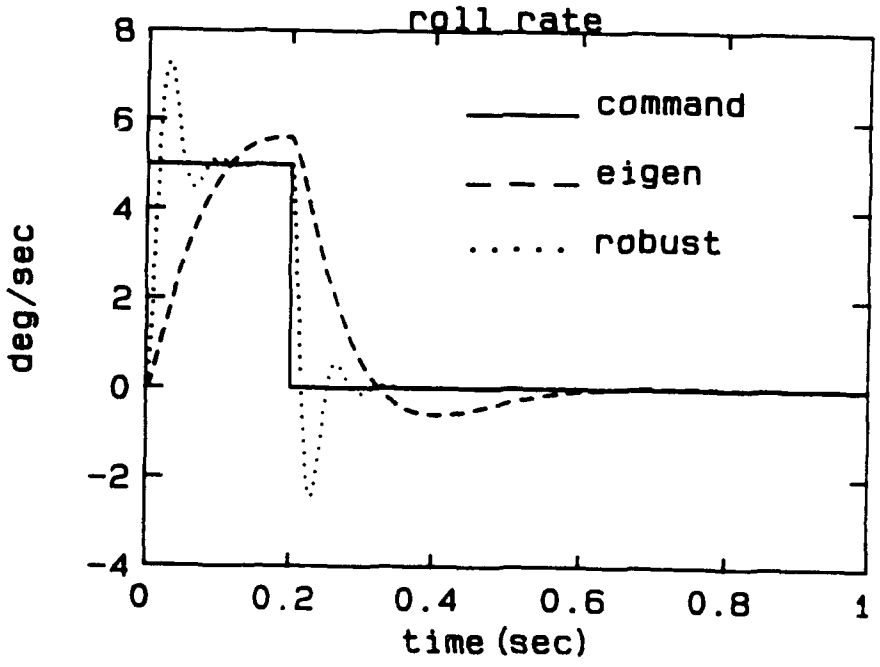


Fig. 7. Roll rate step response (continuous time)

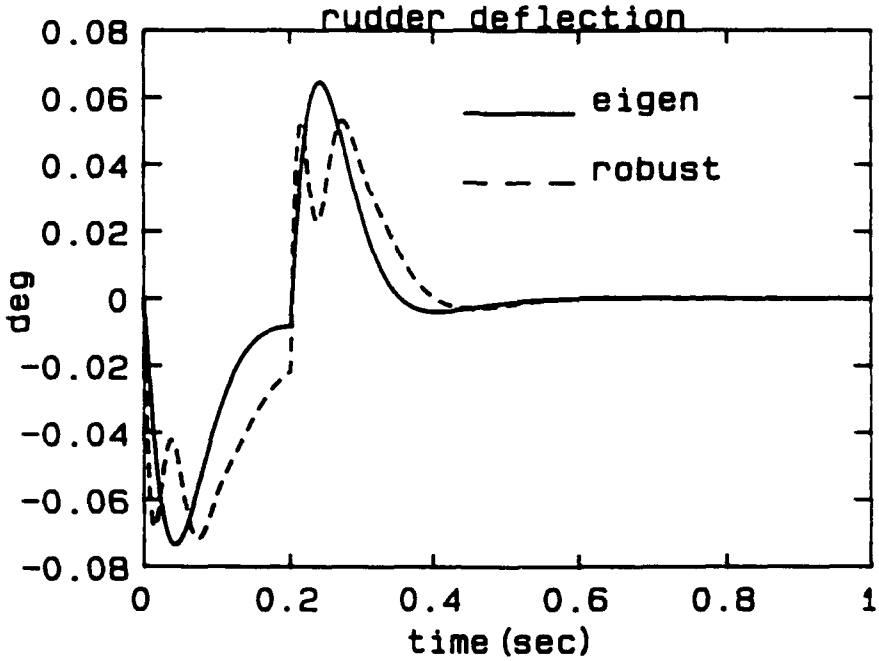


Fig. 8. Rudder deflection for step response (continuous time)

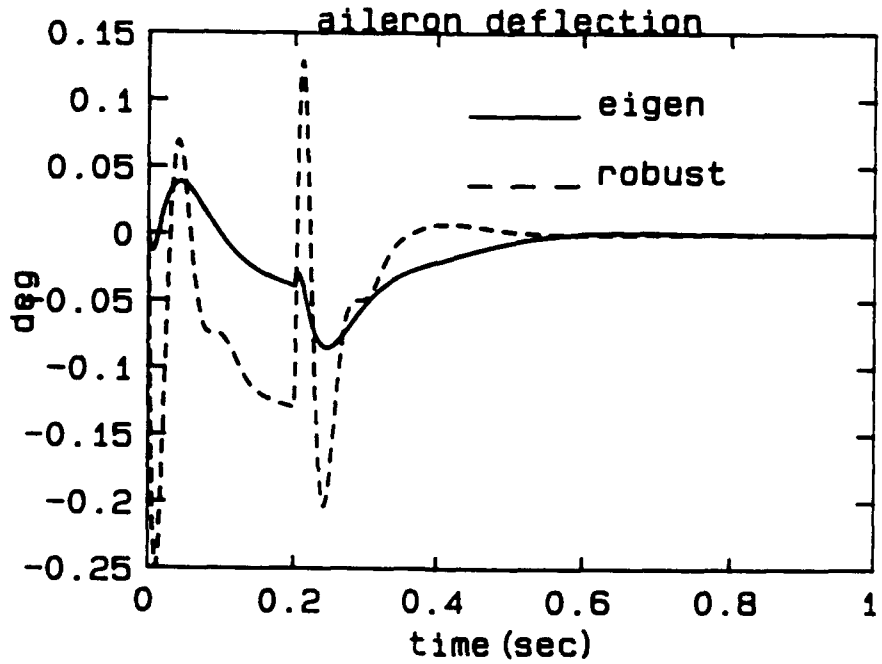


Fig. 9. Aileron deflection for step response (continuous time)

3. ROBUST EIGENSTRUCTURE ASSIGNMENT FOR SAMPLED DATA SYSTEMS: A UNIFIED APPROACH

3.1 Problem Formulation

Consider a nominal linear time-invariant multi-input multi-output system described by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{54}$$

$$y(t) = Cx(t) \tag{55}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the input vector, $y \in \mathbb{R}^r$ is the output vector, and A, B, C, are constant matrices.

The corresponding sampled data system, which is obtained by using Middleton and Goodwin's (1990) delta operator, is shown below.

$$\delta x = A_\delta x + B_\delta u \tag{56}$$

$$y = Cx \tag{57}$$

where

$$A_\delta = \Omega A \tag{58}$$

$$B_\delta = \Omega B \tag{59}$$

$$\Omega = \frac{1}{\Delta} \int_0^\Delta e^{A\tau} d\tau \tag{60}$$

and where

$$\delta = \frac{q-1}{\Delta} \tag{61}$$

and where the shift operator q is defined by

$$qx_k = x_{k+1} \tag{62}$$

and where Δ is the sampling period.

The unified state space model proposed by Middleton and Goodwin (1990) is valid for both the discrete and continuous time cases simultaneously.

This unified model is described by

$$\rho x(t) = A_\rho x(t) + B_\rho u(t) \tag{63}$$

$$y(t) = Cx(t) \tag{64}$$

where

$$A_\rho = \begin{cases} A & \text{in continuous time} \\ A_\delta & \text{in discrete time} \end{cases} \tag{65}$$

$$B_\rho = \begin{cases} B & \text{in continuous time} \\ B_\delta & \text{in discrete time} \end{cases} \tag{66}$$

and where

$$\rho = \begin{cases} \dfrac{d}{dt} & \text{in continuous time} \\ \delta & \text{in discrete time} \end{cases} \tag{67}$$

Suppose that the nominal delta system is subject to linear time-invariant uncertainties in the entries of A_ρ, B_ρ described by dA_ρ and dB_ρ, respectively, where

$$dA_\rho = \begin{cases} dA & \text{in continuous time} \\ dA_\delta & \text{in discrete time} \end{cases} \tag{68}$$

$$dB_\rho = \begin{cases} dB & \text{in continuous time} \\ dB_\delta & \text{in discrete time} \end{cases} \tag{69}$$

and where

$$dA_\delta = \frac{1}{\Delta}\left[e^{(A+dA)\Delta} - e^{A\Delta}\right] \tag{70}$$

$$dB_\delta = \frac{1}{\Delta}\left[\int_0^\Delta e^{(A+dA)\tau}\, d\tau\, (B + dB) - \int_0^\Delta e^{A\tau}\, d\tau\, B\right] \tag{71}$$

Then, the delta system with uncertainty is given by

$$\rho x(t) = A_\rho x(t) + B_\rho u(t) + dA_\rho x(t) + dB_\rho u(t) \tag{72}$$

$$y(t) = Cx(t) \tag{73}$$

Further, suppose that bounds are available on the maximum absolute values of the elements of dA and dB. That is,

$$|da_{ij}| \le (a_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,n \tag{74}$$

$$|db_{ij}| \le (b_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,m \tag{75}$$

Then, the corresponding bounds on the δ system, through Eqs. (70) and (71), are

$$|da_\delta(i,j)| \le [da_\delta(i,j)]_{max};\ i=1,\ldots,n;\ j=1,\ldots,n \tag{76}$$

$$|db_\delta(i,j)| \le [db_\delta(i,j)]_{max};\ i=1,\ldots,n;\ j=1,\ldots,m \tag{77}$$

Define dA_ρ^+ and dB_ρ^+ as the matrices obtained by replacing the entries of dA_ρ and dB_ρ by their absolute values. Also, define $A_{\rho max}$ and $B_{\rho max}$ as the matrices with entries $(a_{ij})_{max}$ and $(b_{ij})_{max}$, respectively in continuous time or with entries $[da_\delta(i,j)]_{max}$ and $[db_\delta(i,j)]_{max}$, respectively in discrete time. Then,

$$\{dA_\rho:\ dA_\rho^+ \le A_{\rho max}\} \tag{78}$$

and

$$\{dB_\rho:\ dB_\rho^+ \le B_{\rho max}\} \tag{79}$$

where

$$A_{\rho max} = \begin{cases} A_{max} & \text{in continuous time} \\ A_{\delta max} & \text{in discrete time} \end{cases} \tag{80}$$

$$B_{\rho max} = \begin{cases} B_{max} & \text{in continuous time} \\ B_{\delta max} & \text{in discrete time} \end{cases} \tag{81}$$

and where

$$A_{\delta max} = \frac{1}{\Delta} \left[e^{(A^+ + A_{max})\Delta} - e^{A^+ \Delta} \right] \tag{82}$$

$$B_{\delta max} = \frac{1}{\Delta} \left[\int_0^{\Delta} e^{(A^+ + A_{max})\tau} d\tau \, (B^+ + B_{max}) - \int_0^{\Delta} e^{A^+ \tau} d\tau \, B^+ \right] \tag{83}$$

and where "$\leq$" is applied element by element to matrices and $A_{max} \in \mathbb{R}_+^{n \times n}$, $B_{max} \in \mathbb{R}_+^{n \times m}$ where $\mathbb{R}_+$ is the set of non-negative numbers.

Consider the constant gain output feedback control law described by

$$u(t) = F_\rho y(t) \tag{84}$$

where

$$F_\rho = \begin{cases} F & \text{in continuous time} \\ F_\delta & \text{in discrete time} \end{cases} \tag{85}$$

Then, the nominal closed loop unified delta system is given by

$$\rho x(t) = A_{\rho c} x(t) \tag{86}$$

where

$$A_{\rho c} = \begin{cases} A + BFC & \text{in continuous time} \\ A_\delta + B_\delta F_\delta C & \text{in discrete time} \end{cases} \tag{87}$$

and the uncertain closed loop unified delta system is given by

$$\rho x(t) = A_{\rho c} x(t) + dA_{\rho c} x(t) \tag{88}$$

where

$$dA_{\rho c} = \begin{cases} dA + dB(FC) & \text{in continuous time} \\ dA_\delta + dB_\delta(F_\delta C) & \text{in discrete time} \end{cases} \tag{89}$$

Finally, the stability robustness problem can be stated as follows: Given a feedback gain matrix $F_\rho \in \mathbb{R}^{m\times r}$ such that the nominal closed loop unified delta system exhibits desirable dynamic performance, determine if the uncertain closed loop unified delta system is asymptotically stable for all time-invariant dA_ρ and dB_ρ described by Eqs.(78) and (79), respectively.

3.2 Robustness Results

Theorem 4: (Delta Eigenvectors)

Consider the continuous time plant given by $\dot{x}=Ax+Bu$ and the sampled data plant given by $\delta x=A_\delta x+B_\delta u$. The i-th eigenvalues of A and A_δ are λ_i and $\gamma_i=\frac{\exp(\lambda_i\Delta)-1}{\Delta}$, respectively. Let M be a non-defective modal matrix and let Λ be a diagonal matrix with the λ_i on the diagonal. Then,

$$M^{-1}A_\delta M = \gamma_i I = \frac{1}{\Delta}(e^{\Lambda\Delta}-I) \text{ if and only if } M^{-1}AM = \lambda_i I = \Lambda. \tag{90}$$

Proof:

Sufficiency: Use the definition of A_δ to obtain

$$M^{-1}A_\delta M = \frac{1}{\Delta} M^{-1}(e^{A\Delta}-I)M \tag{91}$$

Substitute the infinite series for $\exp(A\Delta)$ into Eq.(91) to obtain

$$M^{-1}A_\delta M = \frac{1}{\Delta} M^{-1}[I + A\Delta + \frac{A^2\Delta^2}{2!} + \ldots - I]M \tag{92}$$

$$= \frac{1}{\Delta} [I + M^{-1}AM\Delta + \frac{M^{-1}A^2M\Delta^2}{2!} + \ldots - I] \tag{93}$$

$$= \frac{1}{\Delta} [I + M^{-1}AM\Delta + (M^{-1}AM)^2 \frac{\Delta^2}{2!} + (M^{-1}AM)^3 \frac{\Delta^3}{3!} + \ldots - I] \tag{94}$$

Substitute $M^{-1}AM = \Lambda$ into Eq.(94) to obtain

$$M^{-1}A_\delta M = \frac{1}{\Delta}\left[I + \Lambda\Delta + \Lambda^2\frac{\Delta^2}{2!} + \Lambda^3\frac{\Delta^3}{3!} + \ldots - I\right] \tag{95}$$

$$= \frac{1}{\Delta}\left(e^{\Lambda\Delta}-I\right) \tag{96}$$

Q.E.D.

Necessity:

$$M^{-1}A_\delta M = \frac{1}{\Delta} M^{-1}(e^{A\Delta}-I)M = \frac{1}{\Delta}\left(e^{\Lambda\Delta}-I\right) \tag{97}$$

$$\Rightarrow M^{-1}e^{A\Delta}M = e^{\Lambda\Delta} \tag{98}$$

$$\Rightarrow e^{M^{-1}AM\Delta} = e^{\Lambda\Delta} \tag{99}$$

$$\Rightarrow M^{-1}AM\Delta = \Lambda\Delta \tag{100}$$

$$\Rightarrow M^{-1}AM = \Lambda \tag{101}$$

Q.E.D.

Theorem 5: (Delta Settling Time and Damping Regions)

Consider the plant described in theorem 1 with eigenvalues λ_i in continuous time and eigenvalues γ_i for sampled data operation. The s-plane settling time region described by

$$\mathrm{Re}\ \lambda_i < \sigma \tag{102}$$

maps into the γ-plane region described by

$$|1+\Delta\gamma_i| < e^{\sigma\Delta} \tag{103}$$

and the s-plane damping region described by

$$\cos(\tan^{-1}(\mathrm{Im}\ \lambda_i/(-\mathrm{Re}\ \lambda_i))) > \zeta \tag{104}$$

maps into the γ-plane region described by

$$|1+\Delta\gamma_i| < \exp(-\zeta\phi/(1-\zeta^2)^{1/2}) \tag{105}$$

where $\phi = \arg(1+\Delta\gamma)$.

Proof:

Settling Time Region:

Use $1+\Delta\gamma_i = \exp(\lambda_i\Delta)$ to obtain $|1+\Delta\gamma_i| = \exp(\Delta\cdot\mathrm{Re}\lambda_i)$

Then, $\mathrm{Re}\,\lambda_i < \sigma \Rightarrow |1+\Delta\gamma_i| < \exp(\sigma\Delta)$

Damping Ratio Region:

$$\lambda_i = -\zeta_i\omega_{ni} + j\omega_{di} \tag{106}$$

$$1+\Delta\gamma_i = \exp(\lambda_i\Delta) = \exp(-\zeta_i\omega_{ni}\Delta + j\omega_{di}\Delta) \tag{107}$$

Use $\omega_s\Delta = 2\pi$ and $\omega_{di} = \omega_{ni}(1-\zeta_i^2)^{1/2}$ to obtain

$$1+\Delta\gamma_i = \exp\left[\frac{-\zeta_i 2\pi}{(1-\zeta_i^2)^{1/2}}\frac{\omega_{di}}{\omega_s} + j2\pi\frac{\omega_{di}}{\omega_s}\right] \tag{108}$$

which has a magnitude given by

$$|1+\Delta\gamma_i| = \exp\left[\frac{-\zeta_i 2\pi}{(1-\zeta_i^2)^{1/2}}\frac{\omega_{di}}{\omega_s}\right] = \exp\left[\frac{-\zeta_i\phi_i}{(1-\zeta_i^2)^{1/2}}\right] \tag{109}$$

and a phase angle given by

$$\arg(1+\Delta\gamma_i) = \frac{2\pi\omega_{di}}{\omega_s} \tag{110}$$

$$\text{So } \zeta > \zeta_i \Rightarrow |1+\Delta\gamma_i| < \exp\left[\frac{-\zeta\,\phi}{(1-\zeta^2)^{1/2}}\right] \tag{111}$$

where $\phi = \arg(1+\Delta\gamma)$.

Q.E.D.

Theorem 6: (Delta Eigenstructure Assignment Feedback Gain Matrix)

Suppose $u=F_\delta y$ with $F_\delta \in \mathbb{R}^{m\times r}$ is such that the nominal closed loop system $A_\delta + B_\delta F_\delta C$ is asymptotically stable with a non-defective modal matrix. Let $\Lambda_{\delta r}$ be the $r\times r$ diagonal matrix whose entries are the assignable closed loop eigenvalues and let M_r be the $n\times r$ matrix whose columns are the

corresponding achievable eigenvectors. If a solution exists for

$$(A_\delta + B_\delta F_\delta C)\, M_r = M_r \Lambda_{\delta r} \tag{112}$$

then

$$F_\delta = V_\delta \Sigma_\delta^{-1} U_{\delta 0}^T (M_r \Lambda_{\delta r} - A_\delta M_r) V_r \Sigma_r^{-1} U_{r0}^T \tag{113}$$

where

$$B_\delta = [U_{\delta 0} \quad U_{\delta 1}] \begin{bmatrix} \Sigma_\delta V_\delta^T \\ 0 \end{bmatrix} \tag{114}$$

and

$$CM_r = [U_{r0} \quad U_{r1}] \begin{bmatrix} \Sigma_r V_r^T \\ 0 \end{bmatrix} \tag{115}$$

are the singular value decompositions of B_δ and CM_r, respectively. Furthermore, when $\Delta \to 0$, $F_\delta \to F$ where

$$F = V_B \Sigma_B^{-1} U_{B0}^T (M_r \Lambda_r - A\, M_r) V_r \Sigma_r^{-1} U_{r0}^T \tag{116}$$

is the feedback gain for the continuous time plant which satisifes

$$(A + BFC)\, M_r = M_r \Lambda_r \tag{117}$$

and where

$$B = [U_{B0} \quad U_{B1}] \begin{bmatrix} \Sigma_B V_B^T \\ 0 \end{bmatrix} \tag{118}$$

is a singular value decomposition of B and Λ_r is a diagonal r×r matrix containing the assignable eigenvalues $\lambda_i (i=1,2,\ldots,r)$.

Proof: The r assignable eigenvalues and their corresponding eigenvectors for the sampled data system satisfy the equations described by

$$(A_\delta + B_\delta F_\delta C) v_i = \gamma_i v_i; \quad i=1,2,\ldots,r \tag{119}$$

Combining the r equations from Eq.(119) yields

$$(A_\delta + B_\delta F_\delta C) M_r = M_r \Lambda_{\delta r} \tag{120}$$

Rearrange to obtain

$$B_\delta F_\delta CM_r = M_r\Lambda_{\delta r} - A_\delta M_r \tag{121}$$

Substitute the singular value decompositions of B_δ and CM_r to obtain

$$[U_{\delta 0} \quad U_{\delta 1}]\begin{bmatrix} \Sigma_\delta V_\delta^T \\ 0 \end{bmatrix} F_\delta \; [U_{r0} \quad U_{r1}]\begin{bmatrix} \Sigma_r V_r^T \\ 0 \end{bmatrix} = M_r\Lambda_{\delta r} - A_\delta M_r \tag{122}$$

$$U_{\delta 0}\Sigma_\delta V_\delta^T F_\delta U_{r0}\Sigma_r V_r^T = M_r\Lambda_{\delta r} - A_\delta M_r \tag{123}$$

Taking the required inverses and recalling that $U_{\delta 0}$, U_{r0}, V_δ, and V_r are unitary, we obtain

$$F_\delta = V_\delta\Sigma_\delta^{-1}U_{\delta 0}^T(M_r\Lambda_{\delta r} - A_\delta M_r)V_r\Sigma_r^{-1}U_{r0}^T \tag{124}$$

Next, we consider the limiting behavior of F_δ as $\Delta\to 0$. Middleton and Goodwin (1990) show that as $\Delta\to 0$, $A_\delta\to A$, $B_\delta\to B$, and $\Lambda_{\delta r}\to\Lambda_r$. Thus, $V_\delta\Sigma_\delta^{-1}U_{\delta 0}^T \to V_B\Sigma_B^{-1}U_{B0}^T$, and $F_\delta \to F$.

Q.E.D.

Theorem 7: (Unified Robustness Sufficient Condition)

Suppose that F_ρ is such that the nominal closed loop system described by Eq.(86) is asymptotically stable with a non-defective modal matrix. Then, the uncertain closed loop system given by Eq.(88) is asymptotically stable for dA and dB described by Eqs.(78) and (79), respectively if

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho cmax} \right\} < 1 \tag{125}$$

where

$$f(\gamma_i) = \begin{cases} -\mathrm{Re}(\gamma_i) & \text{continuous time} \\ \dfrac{1}{\Delta}\,[1-(1+\Delta\gamma_i)^+] & \text{discrete time} \end{cases} \tag{126}$$

$$A_{\rho c max} = \begin{cases} A_{max} + B_{max}(FC)^+ & \text{continuous time} \\ A_{\delta max} + B_{\delta max}(F_\delta C)^+ & \text{discrete time} \end{cases} \tag{127}$$

$$A_{\delta max} = \frac{1}{\Delta}\left[e^{(A^+ + A_{max})\Delta} - e^{A^+\Delta}\right]$$

$$B_{\delta max} = \frac{1}{\Delta}\left[\int_0^\Delta e^{(A^+ + A_{max})\tau} d\tau\ (B^+ + B_{max}) - \int_0^\Delta e^{A^+\tau}\, d\tau\ B^+\right] \tag{128}$$

and where γ_i is the i-th eigenvalue of $(A_\rho + B_\rho F_\rho C)$ with v_i and w_i^* the corresponding right and left eigenvectors, respectively; and where $(\cdot)^*$ denotes the complex conjugate transpose.

Proof: The proof starts from showing the following lemma:

Lemma 1: (Delta Uncertainty Bounds)

$$(dA_\delta)^+ \le A_{\delta max} = \frac{1}{\Delta}\left[e^{(A^+ + A_{max})\Delta} - e^{A^+\Delta}\right] \tag{129}$$

$$(dB_\delta)^+ \le B_{\delta max} = \frac{1}{\Delta}\left[\int_0^\Delta e^{(A^+ + A_{max})\tau} d\tau\ (B^+ + B_{max}) - \int_0^\Delta e^{A^+\tau}\, d\tau\ B^+\right] \tag{130}$$

Proof of Lemma 1:

$$(dA_\delta)^+ = \frac{1}{\Delta}\left[e^{(A+dA)\Delta} - e^{A\Delta}\right]^+ = \frac{1}{\Delta}\left\{\left[I + (A+dA)\Delta + (A+dA)^2\frac{\Delta^2}{2!} + (A+dA)^3\frac{\Delta^3}{3!} + \cdots\right] - \left[I + A\Delta + A^2\frac{\Delta^2}{2!} + A^3\frac{\Delta^3}{3!} + \cdots\right]\right\}^+$$

$$= \frac{1}{\Delta}\left\{dA\cdot\Delta + \left[A\cdot dA + dA\cdot A + (dA)^2\right]\frac{\Delta^2}{2!} + \left[A\cdot dA\cdot A + dA\cdot A\cdot dA + dA\cdot A^2 + A^2\cdot dA + (dA)^2\cdot A + A\cdot(dA)^2 + (dA)^3\right]\frac{\Delta^3}{3!} + \cdots\right\}^+$$

$$\leq \frac{1}{\Delta}\left\{A_{max}\cdot\Delta + \left[A^{+}\cdot A_{max} + A_{max}\cdot A^{+} + A_{max}^{2}\right]\frac{\Delta^{2}}{2!} + \left[A^{+}\cdot A_{max}\cdot A^{+} + A_{max}\cdot A^{+}\cdot A_{max}\right.\right.$$

$$\left.\left. + A_{max}\cdot(A^{+})^{2} + (A^{+})^{2}\cdot A_{max} + (A_{max})^{2}\cdot A + A\cdot(A_{max})^{2} + (A_{max})^{3}\right]\frac{\Delta^{3}}{3!} + \cdots\right\}$$

$$= \frac{1}{\Delta}\left\{\left[I + (A^{+} + A_{max})\Delta + (A^{+} + A_{max})^{2}\frac{\Delta^{2}}{2!} + (A^{+} + A_{max})^{3}\frac{\Delta^{3}}{3!} + \cdots\right]\right.$$

$$\left. - \left[I + A^{+}\Delta + (A^{+})^{2}\frac{\Delta^{2}}{2!} + (A^{+})^{3}\frac{\Delta^{3}}{3!} + \cdots\right]\right\}$$

$$= \frac{1}{\Delta}\left[e^{(A^{+}+A_{max})\Delta} - e^{A^{+}\Delta}\right]$$

$$(dB_{\delta})^{+} = \left\{\frac{1}{\Delta}\left[\int_{0}^{\Delta} e^{(A+dA)\tau}\,d\tau\,(B + dB) - \int_{0}^{\Delta} e^{A\tau}\,d\tau\,B\right]\right\}^{+}$$

$$= \frac{1}{\Delta}\left\{\left[I\Delta + (A+dA)\frac{\Delta^{2}}{2!} + (A+dA)^{2}\frac{\Delta^{3}}{3!} + (A+dA)^{3}\frac{\Delta^{4}}{4!} + \cdots\right](B + dB)\right.$$

$$\left. - \left[I\Delta + A\frac{\Delta^{2}}{2!} + A^{2}\frac{\Delta^{3}}{3!} + A^{3}\frac{\Delta^{4}}{4!} + \cdots\right]B\right\}^{+}$$

$$= \frac{1}{\Delta}\left\{\left[I\Delta + A\frac{\Delta^{2}}{2!} + A^{2}\frac{\Delta^{3}}{3!} + A^{3}\frac{\Delta^{4}}{4!} + \cdots\right]dB\right.$$

$$+ \left[dA\cdot\frac{\Delta^{2}}{2!} + \left(A\cdot dA + dA\cdot A + (dA)^{2}\right)\frac{\Delta^{3}}{3!} + \left(A\cdot dA\cdot A + dA\cdot A\cdot dA + dA\cdot A^{2} + A^{2}\cdot dA\right.\right.$$

$$\left.\left. + (dA)^{2}\cdot A + A\cdot(dA)^{2} + (dA)^{3}\right)\frac{\Delta^{4}}{4!} + \cdots\right](B + dB)\Bigg\}^{+}$$

$$\leq \frac{1}{\Delta}\left\{\left[I\Delta + A^{+}\frac{\Delta^{2}}{2!} + (A^{+})^{2}\frac{\Delta^{3}}{3!} + (A^{+})^{3}\frac{\Delta^{4}}{4!} + \cdots\right]B_{max}\right.$$

$$+ \left[A_{max}\cdot\frac{\Delta^{2}}{2!} + \left(A^{+}\cdot A_{max} + A_{max}\cdot A^{+} + (A_{max})^{2}\right)\frac{\Delta^{3}}{3!} + \right.$$

$$\left(A^{+}\cdot A_{max}\cdot A^{+} + A_{max}\cdot A^{+}\cdot A_{max} + A_{max}\cdot(A^{+})^{2} + (A^{+})^{2}\cdot A_{max}\right.$$

$$\left.\left. + (A_{max})^{2}\cdot A^{+} + A^{+}\cdot(A_{max})^{2} + (A_{max})^{3}\right)\frac{\Delta^{4}}{4!} + \cdots\right](B^{+} + B_{max})\Bigg\}$$

$$= \frac{1}{\Delta}\left\{\left[I\Delta + (A^{+} + A_{max})\frac{\Delta^{2}}{2!} + (A^{+} + A_{max})^{2}\frac{\Delta^{3}}{3!} + (A^{+} + A_{max})^{3}\frac{\Delta^{4}}{4!} + \cdots\right](B^{+}\right.$$

$$+ B_{max}) - \left[I\Delta + A^{+}\frac{\Delta^2}{2!} + (A^{+})^2\frac{\Delta^3}{3!} + (A^{+})^3\frac{\Delta^4}{4!} + \cdots \right] B^{+} \Bigg\}$$

$$= \frac{1}{\Delta} \left[\int_0^{\Delta} e^{(A^{+}+A_{max})\tau} \, d\tau \, (B^{+} + B_{max}) - \int_0^{\Delta} e^{A^{+}\tau} \, d\tau \, B^{+} \right]$$

Q.E.D.

The proof of theorem 7 now continues by observing that the uncertain closed loop system may be written as

$$\rho x(t) = A_{\rho c}x(t) + dA_{\rho c}x(t) \tag{131}$$

where

$$A_{\rho c} = \begin{cases} A+BFC & \text{continuous} \\ A_{\delta}+B_{\delta}F_{\delta}C & \text{discrete} \end{cases}$$

and

$$dA_{\rho c} = \begin{cases} dA+dBFC & \text{continuous} \\ dA_{\delta}+dB_{\delta}F_{\delta}C & \text{discrete} \end{cases}$$

which has a solution given by (Middleton & Goodwin)

$$x(t) = E(A_{\rho c}, t)x(0) + \mathop{S}_{0}^{t} E(A_{\rho c}, t-\tau-\Delta)dA_{\rho c}x(\tau)d\tau \tag{132}$$

where

$$E(A_{\rho c}, t) = \begin{cases} e^{A_{\rho c}t} & \text{continuous time} \\ (I+A_{\rho c}\Delta)^{t/\Delta} & \text{discrete time} \end{cases} \tag{133}$$

and

$$S = \begin{cases} \int & \text{continuous time} \\ \Delta\Sigma & \text{discrete time} \end{cases} \tag{134}$$

Next, use the real, positive, diagonal transformation

$$x(t) = D^{-1}z(t) \tag{135}$$

and the properties that (see Middleton and Goodwin 1990)

$$E(DA_{\rho c}D^{-1}t) = DE(A_{\rho c},t)D^{-1} \tag{136}$$

and

$$E(A_{\rho c},t) = ME(\Lambda_{\rho},t)M^{-1} = \sum_{i=1}^{n} v_i w_i^* E(\gamma_i,t) \tag{137}$$

to obtain

$$z(t) = D\sum_{i=1}^{n} v_i w_i^* E(\gamma_i,t)D^{-1}z(0) + \int_0^t D\sum_{i=1}^{n} v_i w_i^* E(\gamma_i,t-\tau-\Delta)dA_{\rho c}(\tau)D^{-1}z(\tau)d\tau \tag{138}$$

where M is a modal matrix of $A_{\rho c}$; γ_i is the i-th eigenvalue of $A_{\rho c}$ given by

$$\gamma_i = \begin{cases} \lambda_i & \text{continuous time} \\ \frac{1}{\Delta}(e^{\lambda_i\Delta}-1) & \text{discrete time} \end{cases} \tag{139}$$

with v_i and w_i^* the corresponding right and left eigenvectors, respectively; Λ_ρ is a diagonal matrix with the γ_i on the diagonal; and $(\cdot)^*$ denotes complex conjugate transpose.

Note that $\|z(t)\| \to 0$ implies that $\|x(t)\| \to 0$. Next, we apply the absolute value operator, denoted by $(\cdot)^+$, to both sides of Eq.(138) where "+" and "≤" are applied element by element to vectors and matrices.

$$z^+(t) \le \left[D\sum_{i=1}^{n} v_i w_i^* E(\gamma_i,t)D^{-1}z(0)\right]^+ + \left[\int_0^t D\sum_{i=1}^{n} v_i w_i^* E(\gamma_i,t-\tau-\Delta)dA_{\rho c}(\tau)D^{-1}z(\tau)d\tau\right]^+ \tag{140}$$

$$\le D\sum_{i=1}^{n} (v_i w_i^*)^+ E^+(\gamma_i,t)D^{-1}z^+(0) + \int_0^t D\sum_{i=1}^{n} (v_i w_i^*)^+ E^+(\gamma_i,t-\tau-\Delta)A_{\rho c\max}(\tau)D^{-1}z^+(\tau)d\tau \tag{141}$$

where

$$A_{\rho c max} = \begin{cases} A_{max} + B_{max}\,(FC)^{+} & \text{continuous time} \\ A_{\delta max} + B_{\delta max}\,(F_{\delta}C)^{+} & \text{discrete time} \end{cases} \tag{142}$$

with $A_{\delta max}$ and $B_{\delta max}$ given by lemma 1; and where

$$E(\gamma_i, t) = \begin{cases} e^{\gamma_i t} & \text{continuous time} \\ (1+\Delta\gamma_i)^{t/\Delta} & \text{discrete time} \end{cases} \tag{143}$$

$$E^{+}(\gamma_i, t) = \begin{cases} e^{Re(\gamma_i)\cdot t} = e^{-\alpha_i t} & \text{continuous time} \\ [(1+\Delta\gamma_i)^{+}]^{t/\Delta} = e^{-\alpha_i k\Delta} & \text{discrete time} \end{cases} \tag{144}$$

and $\alpha_i = -Re(\lambda_i)$.

Next, apply the $\mathcal{S}$ operator to both sides of Eq.(141) to obtain

$$\mathcal{S}_0^t z^{+}(t)dt \le D\sum_{i=1}^{n} (v_i w_i^{*})^{+} \mathcal{S}_0^{\infty} E^{+}(\gamma_i, t)dt D^{-1} z^{+}(0)$$

$$+ \mathcal{S}_{t=0}^{\infty} \mathcal{S}_{\tau=0}^{t} D\sum_{i=1}^{n} (v_i w_i^{*})^{+} E^{+}(\gamma_i, t-\tau-\Delta) A_{\rho c max}(\tau) D^{-1} z^{+}(\tau)d\tau dt \tag{145}$$

where

$$\mathcal{S}_0^t E^{+}(\gamma_i, t)dt = \begin{cases} \int_0^{\infty} e^{Re(\gamma_i)t} dt = \int_0^{\infty} e^{Re(\lambda_i)t} dt = \int_0^{\infty} e^{-\alpha_i t} dt = \dfrac{1}{\alpha_i} & \text{continuous time} \\ \Delta\sum_{k=0}^{\infty} [(1+\Delta\gamma_i)^{+}]^{k} = \Delta\sum_{k=0}^{\infty} e^{-\alpha_i k\Delta} = \dfrac{\Delta}{1-\exp(-\alpha_i\Delta)} & \text{discrete time} \end{cases} \tag{146}$$

Consider the double $\mathcal{S}$ term in Eq.(145) which may be written as

$$\lim_{R\to\infty} \left[\mathcal{S}_{t=0}^{R} \mathcal{S}_{\tau=0}^{t} D\sum_{i=1}^{n} (v_i w_i^{*})^{+} E^{+}(\gamma_i, t-\tau-\Delta) A_{\rho c max}(\tau) D^{-1} z^{+}(\tau)d\tau dt \right] \tag{147}$$

In continuous time, Eq.(147) becomes

$$\lim_{R\to\infty}\left[\int_{t=0}^{R}\int_{\tau=0}^{t} D\sum_{i=1}^{n}(v_i w_i^*)^+ e^{-\alpha_i(t-\tau)}\,[A_{max} + B_{max}(FC)^+]D^{-1}z^+(\tau)d\tau dt\right] \quad (148)$$

which satisfies (see Yu et. al. 1991)

$$\le D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}\,[A_{max} + B_{max}(FC)^+]D^{-1}\int_0^{\infty} z^+(t)dt \quad (149)$$

In discrete time, let $t=k\Delta$ and $\tau=j\Delta$. Then, Eq.(147) becomes

$$\lim_{R\to\infty}\left[\Delta\sum_{k=0}^{R}\Delta\sum_{j=0}^{k} D\sum_{i=1}^{n}(v_i w_i^*)^+\,[(1+\Delta\gamma_i)^+]^{k-j-1}\,A_{\rho cmax}\,D^{-1}\,z^+(j)\right] \quad (150)$$

Upon changing the order of the summations and letting $q=k-j-1$; $q\ge 0$, Eq.(150) becomes

$$\lim_{R\to\infty}\left[\Delta^2\sum_{j=0}^{k}\sum_{q=-j-1}^{k} D\sum_{i=1}^{n}(v_i w_i^*)^+\,[(1+\Delta\gamma_i)^+]^q\,A_{\rho cmax}\,D^{-1}\,z^+(j)\right] \quad (151)$$

$$\le \lim_{R\to\infty}\left[\Delta^2\sum_{j=0}^{k}\sum_{q=0}^{k} D\sum_{i=1}^{n}(v_i w_i^*)^+\,[(1+\Delta\gamma_i)^+]^q\,A_{\rho cmax}\,D^{-1}\,z^+(j)\right] \quad (152)$$

$$\le \lim_{R\to\infty}\left[\Delta^2 D\sum_{i=1}^{n}(v_i w_i^*)^+\sum_{q=0}^{R}[(1+\Delta\gamma_i)^+]^q\,A_{\rho cmax}\,D^{-1}\sum_{j=0}^{k}z^+(j)\right] \quad (153)$$

$$\le \lim_{R\to\infty}\left[\Delta^2 D\sum_{i=1}^{n}(v_i w_i^*)^+\left(\frac{1-[(1+\Delta\gamma_i)^+]^{R+1}}{1-(1+\Delta\gamma_i)^+}\right)A_{\rho cmax}\,D^{-1}\sum_{j=0}^{k}z^+(j)\right] \quad (154)$$

Use Eq.(139) to obtain $(1+\Delta\gamma_i)=\exp(\lambda_i\Delta)$. Hence, $(1+\Delta\gamma_i)^+=\exp(-\alpha_i\Delta)$ and $\lim_{R\to\infty}\left[1-[(1+\Delta\gamma_i)^+]^{R+1}\right] = \lim_{R\to\infty}\left[1-\exp[-\alpha_i\Delta(R+1)]\right] = 1$ because $\alpha_i>0$.

Substitute this result into Eq.(154) to obtain

$$\le \Delta^2 D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{1-(1+\Delta\gamma_i)^+}\,A_{\rho cmax}\,D^{-1}\sum_{j=0}^{\infty}z^+(j) \quad (155)$$

Combine Eqs.(149) and (155) and use Eq.(134) to obtain

$$\int_{t=0}^{\infty} \int_{\tau=0}^{t} D \sum_{i=1}^{n} (v_i w_i^*)^+ E^+(\gamma_i, t-\tau-\Delta) A_{\rho c max} D^{-1} z^+(\tau)\, d\tau\, dt$$

$$\leq D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho c max} D^{-1} \int_0^{\infty} z^+(t)\, dt \tag{156}$$

where

$$f(\gamma_i) = \begin{cases} -\mathrm{Re}(\gamma_i) = -\mathrm{Re}(\lambda_i) = \alpha_i & \text{continuous time} \\ \frac{1}{\Delta}\,[1-(1+\Delta\gamma_i)^+] = \frac{1}{\Delta}\,[1-e^{-\alpha_i \Delta}] & \text{discrete time} \end{cases} \tag{157}$$

Substitute Eq.(156) into Eq.(145) and use Eq.(146) to obtain

$$\int_0^{\infty} z^+(t)\, dt \leq D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} D^{-1} z^+(0) + D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho c max} D^{-1} \int_0^{\infty} z^+(t)\, dt \tag{158}$$

Take norms in Eq,(158) and rearrange to obtain

$$\left\| \int_0^{\infty} z^+(t)\, dt \right\| \leq \frac{\left\| D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} D^{-1} z^+(0) \right\|}{1 - \left\| D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho c max} D^{-1} \right\|} \tag{159}$$

Thus, $\left\| \int_0^{\infty} z^+(t)\, dt \right\| < \infty$ which implies $\int_0^{\infty} \|z^+(t)\|\, dt < \infty$ if

$$\left\| D \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho c max} D^{-1} \right\| < 1 \tag{160}$$

Note that in continuous time, $x(t)$ and $z(t)$ are uniformly continuous on $(0,\infty)$ because of the linearity of the uncertain closed loop plant which together with Eq.(160) implies that $\|z^+(t)\| \to 0$ as $t \to \infty$ (see Hsu and Meyer 1968). This implies that $\|x(t)\| \to 0$ as $t \to \infty$ which proves that the linear uncertain

closed loop plant is asymptotically stable. In discrete time, the same result follows without the need for uniform continuity.

Finally, Perron weightings may be used for the matrix D in Eq.(160) in the same manner as shown by Sobel et. al. (1989) for an earlier robustness result. Thus, to reduce conservatism, Eq.(160) may be replaced by

$$\lambda_{max}\left\{ \sum_{i=1}^{n} \frac{(v_i w_i^*)^+}{f(\gamma_i)} A_{\rho c max} \right\} < 1 \tag{161}$$

where $\lambda_{max}(\cdot)$ of a non-negative matrix denotes the real non-negative eigenvalue $\lambda_{max} \geq 0$ such that $\lambda_{max} \geq |\lambda_i|$ for all eigenvalues λ_i.

Q.E.D.

3.3 Application to EMRAAT Missile

Consider the Extended Medium Range Air to Air Technology (EMRAAT) bank-to-turn missile which is described by Bossi and Langehough (1988). A seventh order model of the yaw/roll dynamics at a 10 degree angle of attack is considered which includes the rigid body dynamics, two first order actuator models, and a yaw rate washout filter. The state vector, control vector, and measurement vector given by $x = [\beta, r, p, p_I, \delta_r, \delta_a, x_7]^T$, $u = [\delta_{rc}, \delta_{ac}]^T$, and $y = [\beta, r_{wo}, p, p_I]^T$, respectively. Here β is the sideslip angle (deg), r is yaw rate (deg/sec), p is roll rate (deg/sec), p_I is integrated roll rate (deg), δ_r is rudder deflection (deg), δ_a is aileron deflection (deg), x_7 is the yaw rate washout filter state, and r_{wo} is washed out yaw rate (deg/sec). The state space matrices A, B, and C are shown below:

$$A = \begin{bmatrix} -.5007 & -.9945 & .1736 & 0 & .109 & .00691 & 0 \\ 16.83 & -.5748 & .01233 & 0 & -132.8 & 27.19 & 0 \\ -322.7 & .3208 & -2.099 & 0 & -1620.0 & -1240.0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -179 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -179 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & -5 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 179 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 179 & 0 \end{bmatrix}^T$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

First, we design an eigenstructure assignment control law by using an orthogonal projection. The delta state space matrices A_δ and B_δ are computed by using the MATLAB™ Delta Toolbox. The sampling period Δ is chosen to be 0.0025 seconds for illustrative purposes. The desired dutch roll and roll mode eigenvalues are achieved exactly because four measurements are available for feedback. The desired dutch roll eigenvectors are chosen to yield a complex mode which is composed of sideslip angle and yaw rate with no coupling to roll rate and integrated roll rate. The desired roll mode eigenvectors are chosen to yield a complex mode which is composed of roll rate and integrated roll rate with no coupling to sideslip angle and yaw rate. Then, the achievable eigenvectors are computed by using the orthogonal projection of the i-th desired eigenvector v_i^d onto the subspace which is spanned by the columns of $(\gamma_i I - A_\delta)^{-1} B_\delta$. The closed loop delta eigenvalues

γ_i; i=1,...,n and the feedback gain matrix F_δ are shown in Table 2. The desired and achievable closed loop eigenvectors are shown in Table 3.

TABLE 2. COMPARISON OF EMRAAT DESIGNS (Δ=0.0025 sec.)

	Closed Loop Eigenvalues[a]	Feedback Gain Matrix
Orthogonal Projection Design	γ_{dr}=-23.65±j16.97 γ_{roll}=-9.98±j10.16 γ_{act} =-130.3 γ_{act} = -103.8 γ_{filter} = -6.97	$\begin{array}{c} \begin{array}{cccc} \beta & r_{wo} & p & p_I \end{array} \\ \begin{bmatrix} -5.43 & .231 & .0043 & .959 \\ 4.42 & -.285 & .0050 & -1.09 \end{bmatrix} \end{array} \begin{array}{c} \\ \delta_{rc} \\ \delta_{ac} \end{array}$
Robust Design	γ_{dr}=-18.26±j14.15 γ_{roll}=-61.39±j99.70 γ_{act} = -112.8 γ_{act} = -48.61 γ_{filter} = -7.42	$\begin{array}{c} \begin{array}{cccc} \beta & r_{wo} & p & p_I \end{array} \\ \begin{bmatrix} -4.16 & .196 & .0154 & 1.27 \\ 2.36 & -.215 & .0757 & 2.16 \end{bmatrix} \end{array} \begin{array}{c} \\ \delta_{rc} \\ \delta_{ac} \end{array}$

[a]Eigenvalues are computed by using feedback gains which are rounded to three significant digits.

TABLE 3 EIGENVECTORS FOR THE EMRAAT DESIGNS[a]

Desired Closed Loop Eigenvectors:

$$\begin{bmatrix} 1 \\ x \\ 0 \\ 0 \\ x \\ x \\ x \end{bmatrix} \pm j \begin{bmatrix} x \\ 1 \\ 0 \\ 0 \\ x \\ x \\ x \end{bmatrix} \qquad \begin{bmatrix} 0 \\ 0 \\ 1 \\ x \\ x \\ x \\ 0 \end{bmatrix} \pm j \begin{bmatrix} 0 \\ 0 \\ x \\ 1 \\ x \\ x \\ 0 \end{bmatrix} \begin{matrix} \beta \\ r \\ p \\ p_I \\ \delta_r \\ \delta_a \\ x_7 \end{matrix}$$

Dutch Roll Mode Roll Mode

Orthogonal Projection Design:

$$\begin{bmatrix} .0187 \\ .9096 \\ -.0006 \\ 0.0000 \\ .1458 \\ -.2391 \\ -.0952 \end{bmatrix} \begin{bmatrix} .0245 \\ .2339 \\ 0.0000 \\ 0.0000 \\ -.0730 \\ .0316 \\ -.1521 \end{bmatrix} \begin{bmatrix} 0.0085 \\ 0.0003 \\ 0.0767 \\ 0.0461 \\ -0.0013 \\ -0.0116 \\ 0.0000 \end{bmatrix} \begin{bmatrix} -.0090 \\ 0.0000 \\ 0.9942 \\ -.0515 \\ .0038 \\ .0240 \\ -0.0001 \end{bmatrix} \begin{matrix} \beta \\ r \\ p \\ p_I \\ \delta_r \\ \delta_a \\ x_7 \end{matrix}$$

Re v_{dr} Im v_{dr} Re v_{roll} Im v_{roll}

Robust Design:

$$\begin{bmatrix} .0263 \\ .0095 \\ -.1571 \\ 0.0050 \\ -.0899 \\ 0.0467 \\ -.1739 \end{bmatrix} \begin{bmatrix} -0.0287 \\ -.9293 \\ -.0051 \\ 0.0043 \\ -.0900 \\ 0.1941 \\ .1542 \end{bmatrix} \begin{bmatrix} -.0003 \\ 0.0001 \\ -.7808 \\ -.0020 \\ -.0139 \\ -.0690 \\ 0.0000 \end{bmatrix} \begin{bmatrix} 0.0014 \\ .0004 \\ -.6196 \\ 0.0077 \\ 0075 \\ 0.0358 \\ 0.0000 \end{bmatrix} \begin{matrix} \beta \\ r \\ p \\ p_I \\ \delta_r \\ \delta_a \\ x_7 \end{matrix}$$

Re v_{dr} Im v_{dr} Re v_{roll} Im v_{roll}

[a]Eigenvalues are computed by using feedback gains which are rounded to three significant digits. Actuator mode eigenvectors are not shown.

In this report, we propose a new robust design which minimizes the integrated roll rate (which is approximately the bank angle) subject to constraints on the time constants of the dutch roll and roll modes, the damping ratios of the dutch roll and roll modes, the aileron and rudder deflection rates, and the new sufficient condition for robust stability. Mathematically, the objective function to be minimized is given by

$$J = \sum_{k=1}^{120} [p_I(k\Delta)]^2 \tag{162}$$

The upper limit on the index k is chosen to include the time interval $k\Delta \in [0,0.3]$ during which most of the transient response occurs. Of course, computation of Eq.(162) requires that a linear simulation be performed during each function evaluation of the optimization. The constraints are shown below where ζ is the damping ratio.

$$|1+\Delta\gamma_{dr}| \in [e^{-50\Delta}, e^{-6\Delta}] \tag{163}$$

$$|1+\Delta\gamma_{roll}| \in [e^{-50\Delta}, e^{-6\Delta}] \tag{164}$$

$$|1+\Delta\gamma_{rudder}| < e^{-50\Delta} \tag{165}$$

$$|1+\Delta\gamma_{aileron}| < e^{-50\Delta} \tag{166}$$

$$1+\Delta\gamma_{dr}| \in \left[\exp\left(\frac{-0.8\phi_{dr}}{[1-(0.8)^2]^{1/2}}\right), \exp\left(\frac{-0.4\phi_{dr}}{[1-(0.4)^2]^{1/2}}\right)\right] \tag{167}$$

where $\phi_{dr} = \arg(1+\Delta\gamma_{dr})$

$$|1+\Delta\gamma_{roll}| \in \left[\exp\left(\frac{-0.8\phi_{roll}}{[1-(0.8)^2]^{1/2}}\right),\ \exp\left(\frac{-0.4\phi_{roll}}{[1-(0.4)^2]^{1/2}}\right)\right] \tag{168}$$

where $\phi_{roll} = \arg(1+\Delta\gamma_{roll})$

$$\frac{|\delta_a[(k+1)\Delta]-\delta_a(k\Delta)|}{\Delta} < 275 \text{ deg/sec} \tag{169}$$

$$\frac{|\delta_r[(k+1)\Delta]-\delta_r(k\Delta)|}{\Delta} < 275 \text{ deg/sec} \tag{170}$$

$$\lambda_{max}\left\{\sum_{i=1}^{7} \frac{(v_i w_i^*)^+}{\alpha_i}\ [A_{\delta max} + B_{\delta max}(F_\delta C)^+]\right\} < 0.999 \tag{171}$$

where for illustrative purposes we have chosen $A_{max} = 0.04 \cdot A^+$ and $B_{max} = 0$ with the exception that the elements of A_{max} which correspond to actuator or washout filter time constants have been set to zero. The matrices $A_{\delta max}$ and $B_{\delta max}$ are computed from Eqs.(82) and (83), respectively.

The actuator deflection rates are computed from the slopes of the time responses of the deflections during the time interval $k\Delta \in [0,0.03]$. This interval is chosen because the slopes of the deflections are largest during this time interval. Mathematically,

$$\frac{|\delta_a[(k+1)\Delta]-\delta_a(k\Delta)|}{\Delta} < \frac{\max|\delta_a(m\Delta)|}{m\Delta} \tag{172}$$

$$\frac{|\delta_r[(k+1)\Delta]-\delta_r(k\Delta)|}{\Delta} < \frac{\max|\delta_r(m\Delta)|}{m\Delta} \tag{173}$$

where m=0,1,...,12. The maximum deflection rates chosen for the constraints are well within the expected 400 deg/sec limit for the advanced state of the art electromechanical actuator described by Langehough and Simons (1988).

The parameter vector contains the quantities which may be varied by the optimization. This twelve dimensional vector includes Re γ_{dr}, Im γ_{dr}, Re γ_{roll}, Im γ_{roll}, Re $z_1(1)$, Re $z_1(2)$, Im $z_1(1)$, Im $z_1(2)$, Re $z_3(1)$, Re $z_3(2)$, Im $z_3(1)$, Im $z_3(2)$. Here, the two dimensional complex vectors z_i contain the free eigenvector parameters. That is, the i-th eigenvector v_i may be written as

$$v_i = L_i z_i \tag{174}$$

where the columns of $L_i=(\gamma_i I-A_\delta)^{-1}B_\delta$ are a basis for the subspace in which the i-th eigenvector must reside. Thus, the free parameters are the vectors z_i rather than the eigenvectors v_i.

The optimization is performed by using subroutine *constr* from the MATLAB™ Optimization Toolbox (Grace 1990) and subroutine *delsim* from the MATLAB™ Delta Toolbox on a 486™ 25MHz personal computer. The optimization is initialized with the orthogonal projection design which yields an initial value of 9.1360 for the objective function of Eq.(162) and a value of 2.2014 for the left hand side (LHS) of the robustness condition of Eq.(10). The optimization is complete after 3640 function evaluations and yields an optimal objective function of 0.0895 and a value of 0.999 for the LHS of the robustness condition of Eq. (10). We observe from Table 1 that the dutch roll mode is dominant in the robust design whereas the roll mode was chosen to be dominant in

the orthogonal projection design.

The time histories of sideslip angle, yaw rate, roll rate, integrated roll rate, rudder deflection, and aileron deflection to a one degree initial sideslip are shown in figures 10, 11, 12, 13, 14, and 15, respectively. We observe a significant improvement in the integrated roll rate response (which is desired to be zero) when compared to the initial orthogonal projection eigenstructure assignment design. The initial design has a minimum $p_I(t)$ of -.464 deg but the new design of this paper has a minimum $p_I(t)$ of -.103 deg which is an improvement of approximately 78%. We note that this improved response is obtained with both smaller aileron and rudder deflections. Furthermore, it is interesting that the aileron in the initial design exhibits an initial positive deflection of approximately three degrees before becoming negative, whereas this positive initial aileron deflection is only approximately one degree in the robust design.

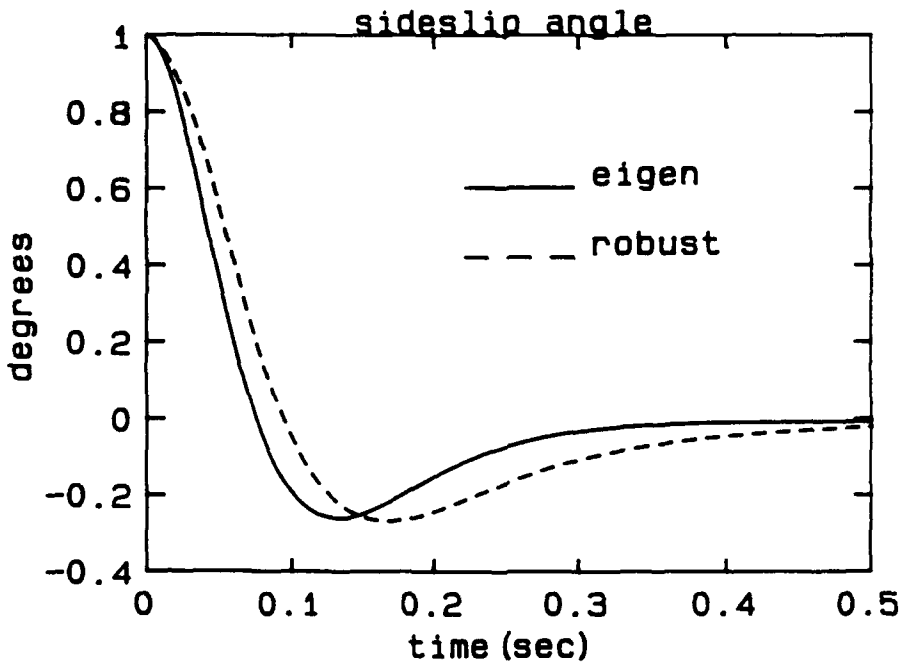


Fig. 10. Sideslip angle (delta model; $\beta(0)$=1 deg)

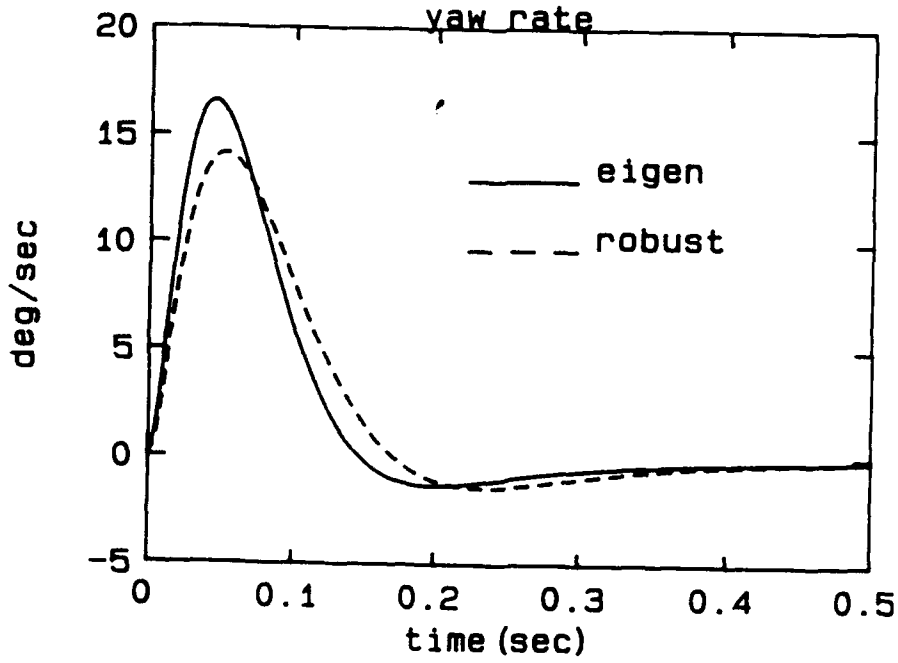


Fig. 11. Yaw rate (delta model; $\beta(0)$=1 deg)

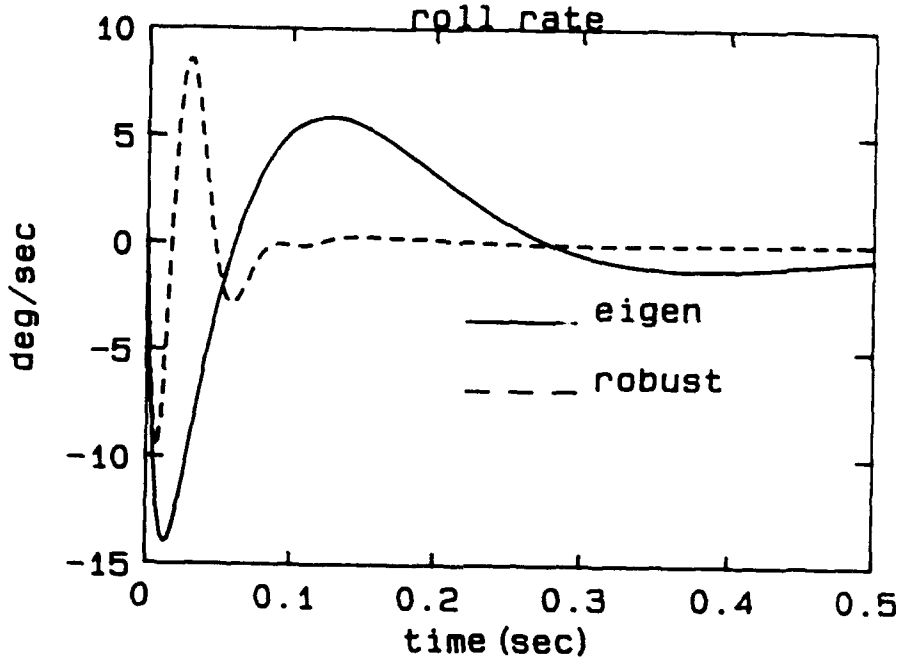


Fig. 12. Roll rate (delta model; $\beta(0)=1$ deg)

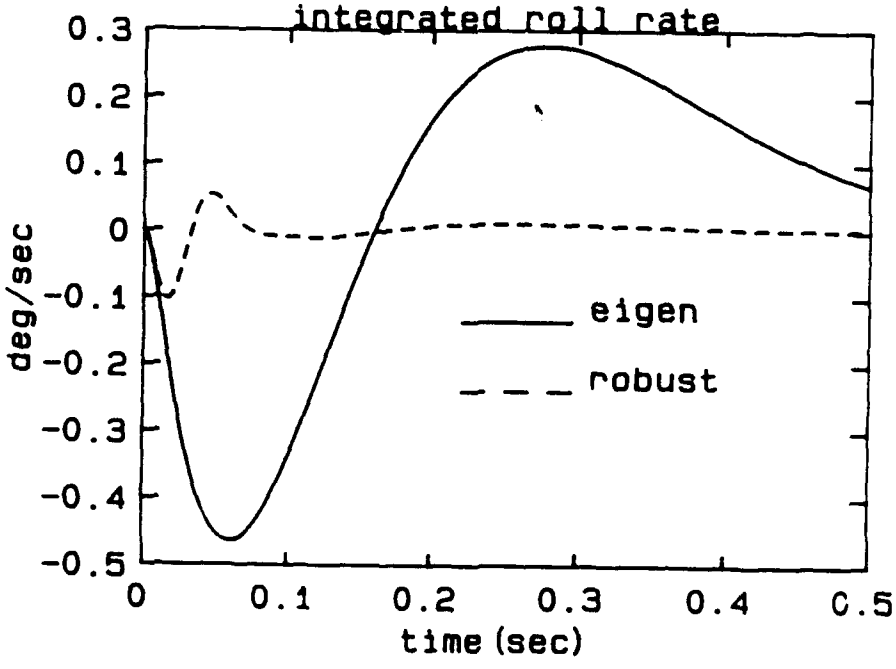


Fig. 13. Integrated roll rate (delta model; $\beta(0)=1$ deg)

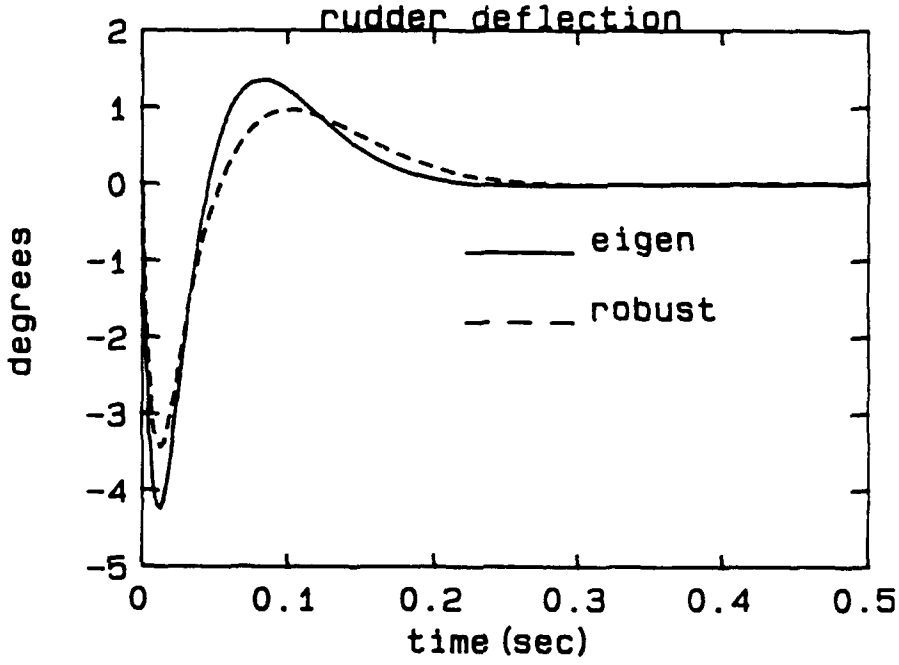


Fig. 14. Rudder deflection (delta model; $\beta(0)=1$ deg)

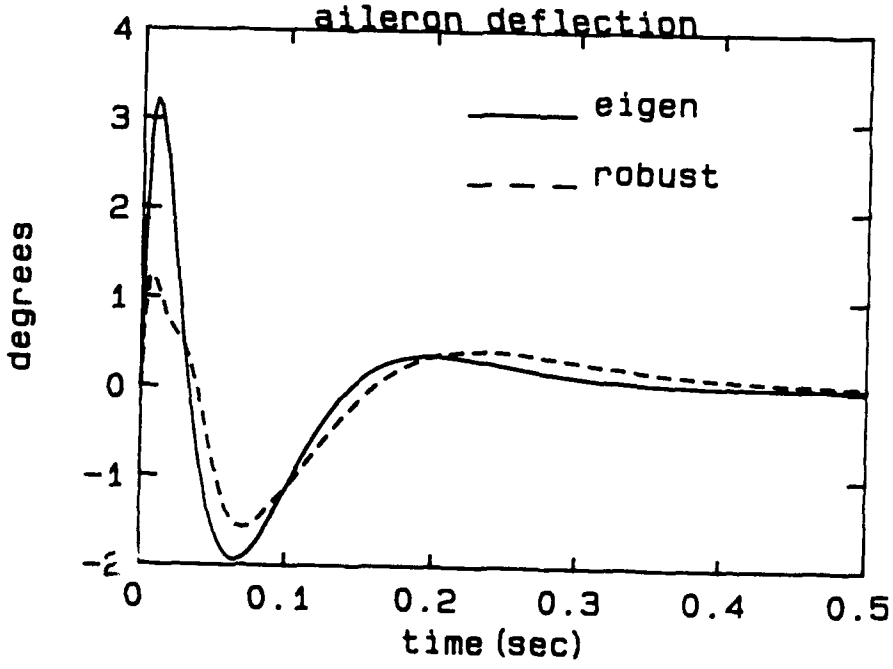


Fig. 15. Aileron deflection (delta model; $\beta(0)=1$ deg)

Next, we compute the time responses due to a "1-cosine" sideslip gust as described in MIL-F-8785-C. The state equations for the lateral dynamics are shown by McRuer et. al. (1973) to be given by

$$\dot{\beta} = Y_v\beta + (g/U_0)\phi - r + (Y_{\delta_a}/U_0)\delta_a + (Y_{\delta_r}/U_0)\delta_r - Y_v\beta_g \tag{175}$$

$$\dot{p} = L'_\beta\beta + L'_p p + L'_r r + L'_{\delta_a}\delta_a + L'_{\delta_r}\delta_r - L'_\beta\beta_g - (L'_r)_g\,\dot{\beta}_g \tag{176}$$

$$\dot{r} = N'_\beta\beta + N'_p p + N'_r r + N'_{\delta_a}\delta_a + N'_{\delta_r}\delta_r - N'_\beta\beta_g - (N'_r)_g\,\dot{\beta}_g \tag{177}$$

where Y_v, Y_v/U_0, Y_{δ_a}/U_0, Y_{δ_r}/U_0, L'_β, L'_p, L'_r, L'_{δ_a}, L'_{δ_r}, N'_β, N'_p, N'_r, N'_{δ_a}, and N'_{δ_r} can be obtained from the state space matrix A and where (see McRuer et. al. 1973)

$$(N'_r)_g = N_r\left[1 - \frac{I_{xz}^2}{I_x I_z}\right]^{-1} \tag{178}$$

$$(L'_r)_g = \left(\frac{I_{xz}}{I_x}\right) N_r\left[1 - \frac{I_{xz}^2}{I_x I_z}\right]^{-1} \tag{179}$$

$$N_r = N'_r - \frac{I_{xz}}{I_z} L'_r \tag{180}$$

For the flight condition of the EMRAAT missile considered in this paper, the parameters N'_r and L'_r are -0.5748 and 0.3208, respectively. The inertias corresponding to a full fuel condition are I_x = 11451, I_z = 456282, and I_{xz} = -1189. Using these values, the gust derivative

coefficients are computed to be $(N_r')_g = -0.5742$ and $(L_r')_g = 0.05962$. The gust is defined as shown in MIL-F-8785-C and is described by

$$\beta_g = \begin{cases} 0 & t<0 \\ 0.5(1-\cos 24\pi t) & 0 \le t \le 1/24 \\ 1 & t > 1/24 \end{cases} \tag{181}$$

where the natural frequency of the open loop complex eigenvalue pair is 24.04 rad/sec.

The time histories of sideslip angle, yaw rate, roll rate, integrated roll rate, rudder deflection, and aileron deflection to the "1-cosine" sideslip gust are shown in figures 16, 17, 18, 19, 20, and 21, respectively. We observe a significant improvement in the integrated roll rate response (which is desired to be zero) when compared to the initial orthogonal projection eigenstructure assignment design. The initial design has a maximum $p_I(t)$ of 18.87 deg but the new design of this paper has a maximum $p_I(t)$ of 0.6990 deg which is an improvement of approximately 96%. We note that this improved response is obtained with both smaller aileron and rudder deflections.

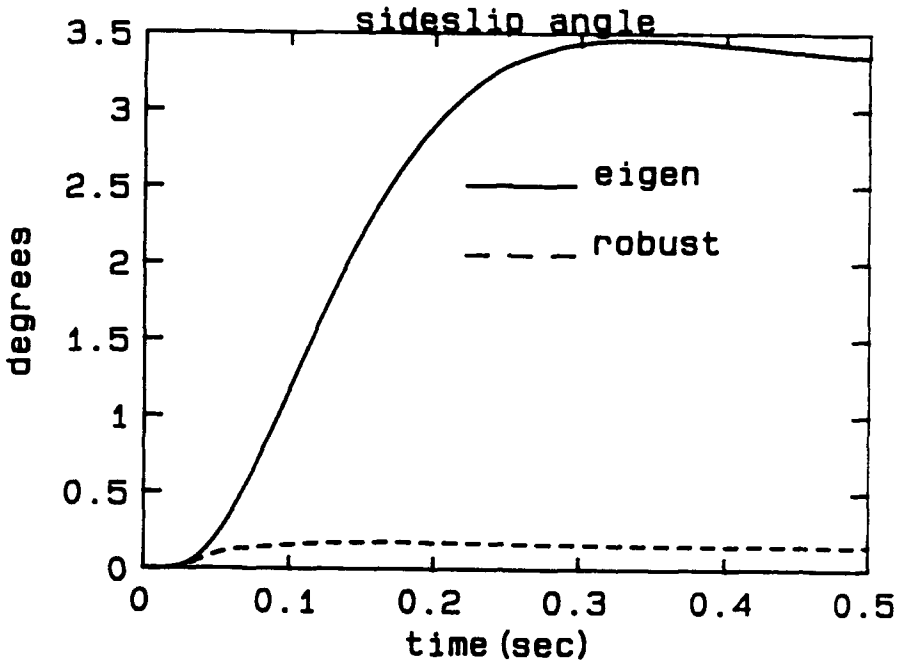


Fig. 16. Sideslip angle (delta model; β_g = 1-cosine)

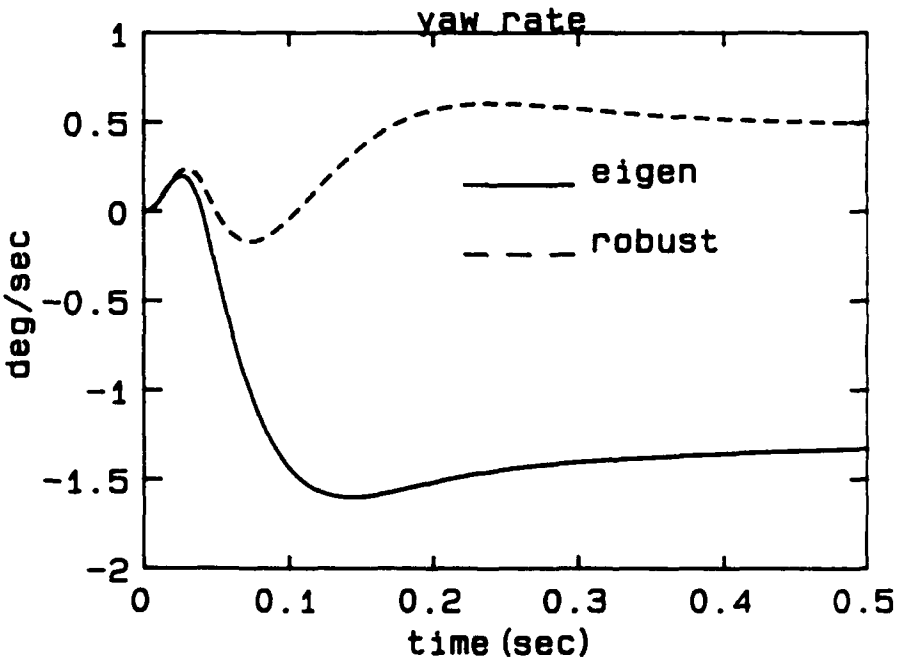


Fig. 17. Yaw rate (delta model; β_g = 1-cosine)

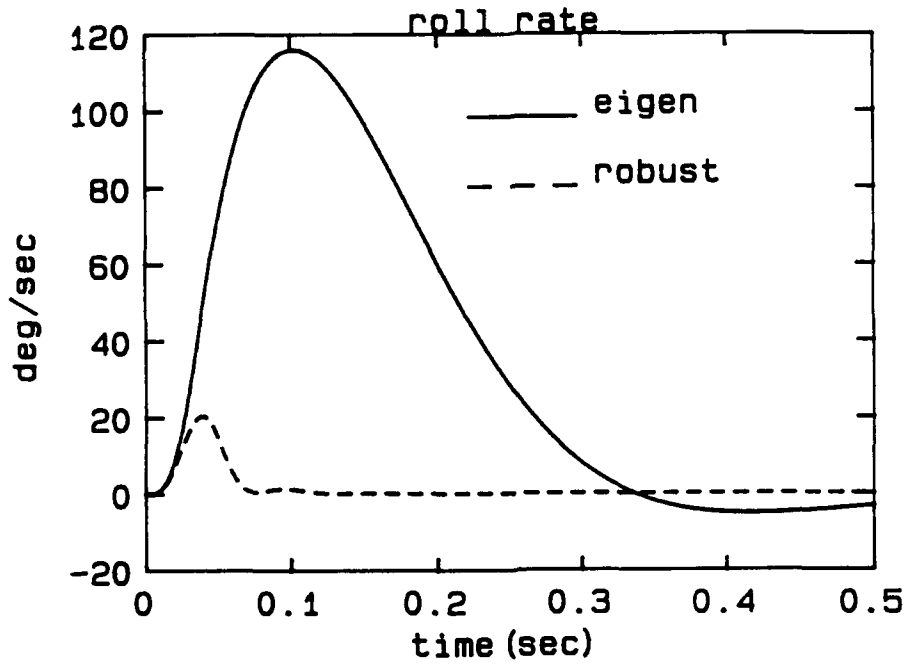


Fig. 18. Roll rate (delta model; β_g = 1-cosine)

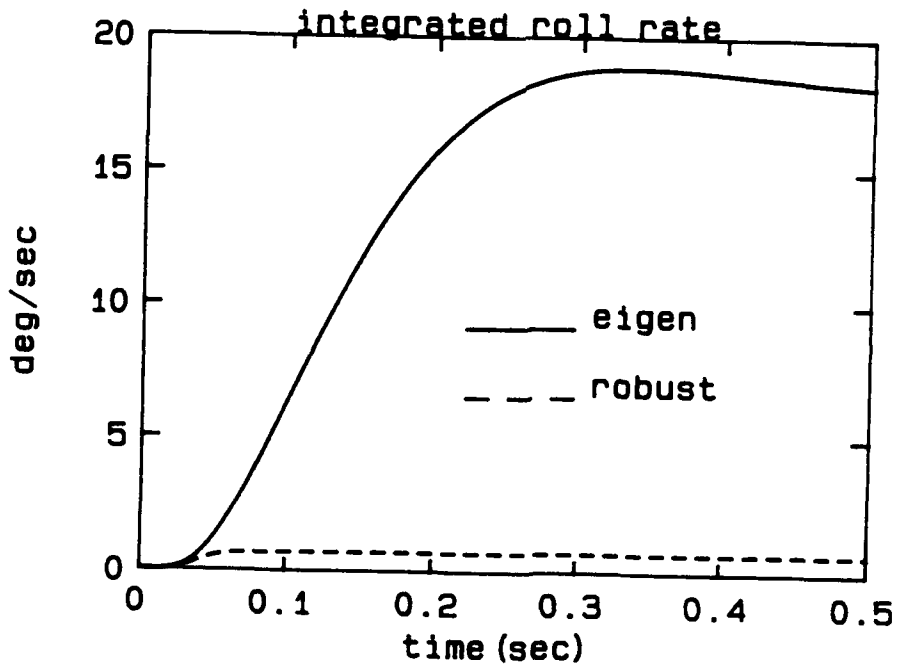


Fig. 19. Integrated roll rate (delta model; β_g = 1-cosine)

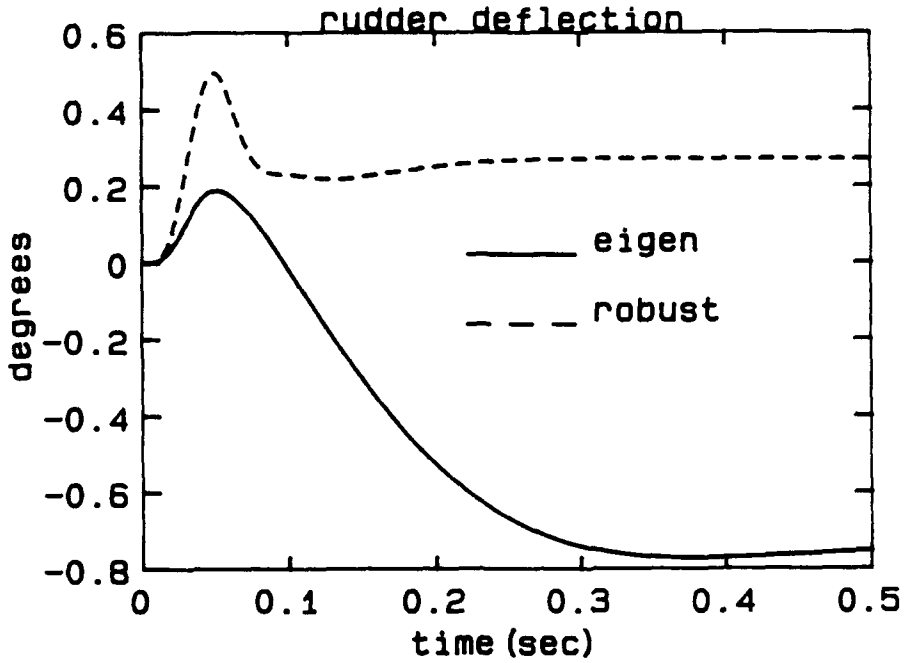


Fig. 20. Rudder deflection (delta model; β_g = 1-cosine)

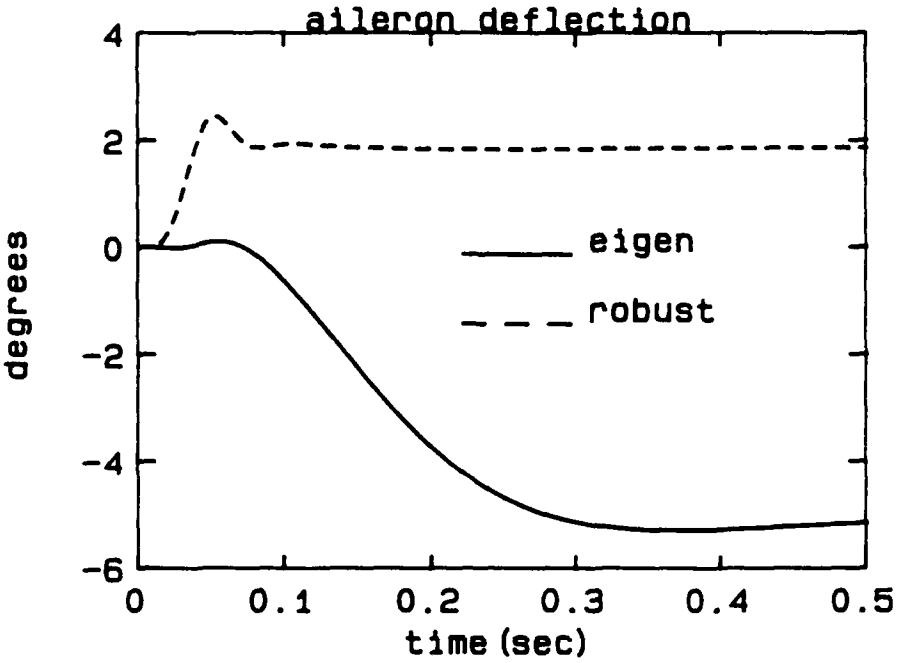


Fig. 21. Aileron deflection (delta model; β_g = 1-cosine)

4. ROBUST EIGENSTRUCTURE ASSIGNMENT FOR TIME DELAY SYSTEMS

Consider a nominal linear time invariant system with time delay described by

$$\dot{x}(t) = Ax(t) + A_0x(t-\tau) + Bu(t) \tag{182}$$

$$y(t) = Cx(t) \tag{183}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}^m$ is the input vector, $y \in \mathbb{R}^r$ is the output vector, and A, B, C, and A_0 are constant matrices.

Suppose that the nominal time delay system is subject to linear time-varying parametric uncertainty in the entries of A, B, C, and A_0 described by $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$, and $\Delta A_0(t)$, respectively. We shall assume that the entries of $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$, and $\Delta A_0(t)$ are uniformly continuous for $t \in (-\infty,\infty)$. Then, the uncertain time delay system is given by

$$\dot{x}(t) = [A+\Delta A(t)]x(t) + [A_0+\Delta A_0(t)]x(t-\tau) + [B+\Delta B(t)]u(t) \tag{184}$$

$$y(t)=[C+\Delta C(t)]x(t) \tag{185}$$

Further, suppose that bounds are available on the absolute values of the maximum variations in the elements of $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$, and $\Delta A_0(t)$. That is,

$$|\Delta a_{ij}(t)| \le (a_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,n \tag{186}$$

$$|\Delta b_{ij}(t)| \le (b_{ij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,m \tag{187}$$

$$|\Delta c_{ij}(t)| \le (c_{ij})_{max};\ i=1,\ldots,r;\ j=1,\ldots,n \tag{188}$$

$$|\Delta a_{oij}(t)| \le (a_{oij})_{max};\ i=1,\ldots,n;\ j=1,\ldots,n \tag{189}$$

Define $\Delta A^+(t)$, $\Delta B^+(t)$, $\Delta C^+(t)$, and $\Delta A_0^+(t)$ as the matrices obtained by replacing the entries of $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$, and $\Delta A_0(t)$ by their absolute values. Also, define A_m, B_m, C_m, and A_{mo} as the matrices with entries $(a_{ij})_{max}$, $(b_{ij})_{max}$, $(c_{ij})_{max}$, and $(a_{oij})_{max}$, respectively. Then,

$$\{\Delta A(t):\ \Delta A^+(t) \le A_m\} \tag{190}$$

$$\{\Delta B(t):\ \Delta B^+(t) \le B_m\} \tag{191}$$

$$\{\Delta C(t):\ \Delta C^+(t) \le C_m\} \tag{192}$$

$$\{\Delta A_0(t):\ \Delta A_0^+(t) \le A_{mo}\} \tag{193}$$

where "$\le$" is applied element by element to matrices and $A_m \in \mathbb{R}_+^{n\times n}$, $B_m \in \mathbb{R}_+^{n\times m}$, $C_m \in \mathbb{R}_+^{r\times n}$, $A_{mo} \in \mathbb{R}_+^{n\times n}$ where $\mathbb{R}_+$ is the set of non-negative numbers.

Consider the constant gain output feedback control law described by

$$u(t) = Fy(t) \tag{194}$$

Then, the nominal closed loop system is given by

$$\dot{x}(t) = (A+BFC)x(t) + A_0x(t-\tau) \tag{195}$$

and the uncertain closed loop system is given by

$$\begin{aligned}\dot{x}(t)&=(A+\Delta A)x(t)+(A_0+\Delta A_0)x(t-\tau)+(B+\Delta B)F(C+\Delta C)x(t)\\&=(A+BFC)x(t)+(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)x(t)+(A_0+\Delta A_0)x(t-\tau)\end{aligned} \tag{196}$$

Finally, the stability robustness problem can be stated as follows: Given a feedback gain matrix $F \in \mathbb{R}^{m\times r}$ such that the nominal closed loop system exhibits desirable dynamic performance, determine if the

uncertain closed loop system is asymptotically stable for all $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$, and $\Delta A_0(t)$ described by Eqs.(190)-(193).

We now present the first robust stability theorem which extends the result of Sobel et. al. (1989) to time delay systems.

Theorem 8: Suppose that F is such that the nominal closed loop system without time delay described by $\dot{x}(t)=(A+BFC)x(t)$ is asymptotically stable with a non-defective modal matrix. Then, the uncertain closed loop system with time delay given by Eq.(196) is asymptotically stable for all $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$ and $\Delta A_0(t)$ described by Eqs.(190)-(193) if

$$\alpha \geq \|(M^{-1})^{+}[A_m+B_m(FC)^{+}+(BF)^{+}C_m+B_mF^{+}C_m]M^{+}\|+\exp(\alpha\tau)\|(M^{-1})^{+}(A_0^{+}+A_{m0})M^{+}\| \qquad (197)$$

where

$$\alpha = -\max_i \operatorname{Re}[\lambda_i(A+BFC)] \qquad (198)$$

and where M is a modal matrix of (A+BFC).

Proof: $\dot{x}(t)=(A+\Delta A)x(t)+(A_0+\Delta A_0)x(t-\tau)+(B+\Delta B)u(t)$ (199)

$y(t)=(C+\Delta C)x(t)$ (200)

$u(t)=Fy(t)$ (201)

$$\dot{x}(t)=(A+\Delta A)x(t)+(A_0+\Delta A_0)x(t-\tau)+(B+\Delta B)F(C+\Delta C)x(t)$$

$$=(A+BFC)x(t)+(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)x(t)+(A_0+\Delta A_0)x(t-\tau) \qquad (202)$$

We begin with two preliminary lemmas.

Lemma 2: $\|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|_2 \leq \|(M^{-1})^{+}\{A_m+B_m(FC)^{+}+(BF)^{+}C_m+B_mF^{+}C_m\}^{+}M^{+}\|$

Proof: Use the result given by Kouvaritakis and Latchman (1985) that

for any matrix $A \in \mathbb{C}^{n\times m}$

$$\|A\|_2 \le \|A^+\|_2$$

Thus

$$\|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|_2 \le \|\{M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\}^+\|_2$$

$$=\sup\frac{\|\{M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\}^+x\|_2}{\|x\|_2} = \sup\frac{\|\{M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\}^+x^+\|_2}{\|x^+\|_2}$$

$$=\sup\frac{\|(M^{-1})^+(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)^+M^+x^+\|_2}{\|x^+\|_2}$$

$$=\sup\frac{\|(M^{-1})^+\{A_m+B_m(FC)^++(BF)^+C_m+B_mF^+C_m\}M^+x^+\|_2}{\|x^+\|_2}$$

$$=\|(M^{-1})^+\{A_m+B_m(FC)^++(BF)^+C_m+B_mF^+C_m\}^+M^+\| \tag{203}$$

Q.E.D.

Lemma 3: $\|M^{-1}(A_0+\Delta A_0)M\|_2 \le \|(M^{-1})^+(A_0^++A_{mo})M^+\|_2$

Proof: $\|M^{-1}(A_0+\Delta A_0)M\|_2 \le \|\{M^{-1}(A_0+\Delta A_0)M\}^+\|_2$

$$=\sup\frac{\|\{M^{-1}(A_0+\Delta A_0)M\}^+x\|_2}{\|x\|_2}=\sup\frac{\|\{M^{-1}(A_0+\Delta A_0)M\}^+x^+\|_2}{\|x^+\|_2}$$

$$=\sup\frac{\|(M^{-1})^+(A_0+\Delta A_0)^+M^+x^+\|_2}{\|x^+\|_2}=\sup\frac{\|(M^{-1})^+(A_0^++A_{mo})M^+x^+\|_2}{\|x^+\|_2}$$

$$=\|(M^{-1})^+(A_0^++A_{mo})M^+\|_2 \tag{204}$$

Q.E.D.

Returning to the proof of theorem 8, we let $x(t)=Mz(t)$

then $z(t)=M^{-1}x(t)$, $\Lambda=M^{-1}(A+BFC)M$

$$z(t)=\exp(\Lambda t)z(0) + \int_0^t \exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)ds$$
$$+ \int_0^t \exp\{\Lambda(t-s)\}M^{-1}(A_0+\Delta A_0)Mz(s-\tau)ds \tag{205}$$

Since $z(t)$ is causal, the second integral should start from τ,

$$z(t)=\exp(\Lambda t)z(0) + \int_0^t \exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)ds$$
$$+ \int_\tau^t \exp\{\Lambda(t-s)\}M^{-1}(A_0+\Delta A_0)Mz(s-\tau)ds \tag{206}$$

Let $u=s-\tau$, i.e. $s=u+\tau$, $ds=du$, then

$$z(t)=\exp(\Lambda t)z(0) + \int_0^t \exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)ds$$
$$+ \int_0^{t-\tau} \exp\{\Lambda(t-u-\tau)\}M^{-1}(A_0+\Delta A_0)Mz(u)du \tag{207}$$

Take the 2-norm on both sides of the above equation to obtain

$$\|z(t)\|=\|\exp(\Lambda t)z(0) + \int_0^t \exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)ds$$
$$+ \int_0^{t-\tau} \exp\{\Lambda(t-u-\tau)\}M^{-1}(A_0+\Delta A_0)Mz(u)du\| \tag{208}$$

$$\le \|\exp(\Lambda t)z(0)\| + \|\int_0^t \exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)ds\|$$
$$+ \|\int_0^{t-\tau} \exp\{\Lambda(t-u-\tau)\}M^{-1}(A_0+\Delta A_0)Mz(u)du\| \tag{209}$$

$$\le \|\exp(\Lambda t)z(0)\| + \int_0^t \|\exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)\|ds$$
$$+\int_0^{t-\tau} \|\exp\{\Lambda(t-u-\tau)\}M^{-1}(A_0+\Delta A_0)Mz(u)\|du \tag{210}$$

$$\le \|\exp(\Lambda t)z(0)\| + \int_0^t \|\exp\{\Lambda(t-s)\}M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)Mz(s)\|ds$$
$$+ \int_0^t \|\exp\{\Lambda(t-u-\tau)\}M^{-1}(A_0+\Delta A_0)Mz(u)\|du \tag{211}$$

$$\le \|\exp(\Lambda t)\|\cdot\|z(0)\|$$
$$+ \int_0^t \|\exp\{\Lambda(t-s)\}\|\cdot\|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|\cdot\|z(s)\|ds$$
$$+ \int_0^t \|\exp\{\Lambda(t-s-\tau)\}\|\cdot\|M^{-1}(A_0+\Delta A_0)M\|\cdot\|z(s)\|ds \tag{212}$$

Using $\|\exp(\Lambda t)\|\le\exp(-\alpha t)$, where $\alpha=-\max[\lambda_i(A+BFC)]$, we obtain

$$\|z(t)\| \le \exp(-\alpha t)\cdot\|z(0)\|$$
$$+ \int_0^t \exp\{-\alpha t+\alpha s\}\cdot\|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|\cdot\|z(s)\|ds$$

$$+ \int_0^t \exp\{-\alpha t+\alpha s+\alpha\tau\}\cdot\|M^{-1}(A_0+\Delta A_0)M\|\cdot\|z(s)\|ds \qquad (213)$$

$$\|z(t)\|\exp(\alpha t) \le \|z(0)\|$$

$$+ \int_0^t \|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|\cdot\|z(s)\|\exp(\alpha s)ds$$

$$+ \int_0^t \exp(\alpha\tau)\|M^{-1}(A_0+\Delta A_0)M\|\cdot\|z(s)\|\exp(\alpha s)ds \qquad (214)$$

Using the Gronwall lemma to obtain

$$\|z(t)\|\exp(\alpha t)\le \|z(0)\|$$

$$\cdot\exp\{\int_0^t[\|M^{-1}(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)M\|+\exp(\alpha\tau)\|M^{-1}(A_0+\Delta A_0)M\|]ds\} \qquad (215)$$

Use Lemmas 2 and 3 to obtain

$$\|z(t)\|\exp(\alpha t)\le \|z(0)\|\cdot\exp\{[\|(M^{-1})^+A_{cm}M^+\|+\exp(\alpha\tau)\|(M^{-1})^+A_{cmo}M^+\|]t\} \qquad (216)$$

where $A_{cm} = A_m+B_m(FC)^++(BF)^+C_m+B_mF^+C_m$

and $A_{cmo} = A_0+A_{mo}$

or $\|z(t)\| \le \|z(0)\|\cdot\exp\left[\left(\{\|(M^{-1})^+A_{cm}M^+\|+\exp(\alpha\tau)\|(M^{-1})^+A_{cmo}M^+\|\}-\alpha\right)t\right]$

Thus $\|z(t)\| \to 0$ if

$$\alpha \ge \|(M^{-1})^+[A_m+B_m(FC)^++(BF)^+C_m+B_mF^+C_m]M^+\|+\exp(\alpha\tau)\|(M^{-1})^+(A_0^++A_{mo})M^+\|$$

Q.E.D.

We next present the second stability robustness theorem which extends the result of theorem 2 (section 2.1) to time delay systems.

Theorem 9: Suppose that F is such that the nominal closed loop system without time delay described by $\dot{x}(t)=(A+BFC)x(t)$ is asymptotically stable with a non-defective modal matrix. Then, the uncertain closed loop system with time delay given by Eq.(196) is asymptotically stable for all $\Delta A(t)$, $\Delta B(t)$, $\Delta C(t)$ and $\Delta A_0(t)$ described by Eqs. (190)-(193)

if

$$\lambda_{max}\left\{\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax}+\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}e^{\alpha_i\tau}A_{comax}\right\} < 1 \tag{217}$$

where

$$\alpha_i = -\mathrm{Re}[\lambda_i(A+BFC)], \tag{218}$$

and where λ_i is the i-th eigenvalue of (A+BFC) with v_i and w_i the corresponding right and left eigenvectors, respectively; and where $(\cdot)^*$ denotes the complex conjugate transpose.

Proof: The uncertain closed loop plant may be written as

$$\begin{aligned}\dot{x}(t)&=(A+\Delta A)x(t)+(A_0+\Delta A_0)x(t-\tau)+(B+\Delta B)F(C+\Delta C)x(t)\\&=(A+BFC)x(t)+(\Delta A+\Delta BFC+BF\Delta C+\Delta BF\Delta C)x(t)+(A_0+\Delta A_0)x(t-\tau)\\&=A_c x(t)+\Delta A_c x(t)+\Delta A_{co}x(t-\tau)\end{aligned} \tag{219}$$

where

$$A_c = A + BFC \tag{220}$$

$$\Delta A_c = \Delta A(t) + \Delta B(t)FC+BF\Delta C(t)+\Delta B(t)F\Delta C(t) \tag{221}$$

$$\Delta A_{co} = A_0 + \Delta A_0(t) \tag{222}$$

which has a solution given by

$$\begin{aligned}x(t) &= \exp(A_c t)x(0)\\&+\int_0^t \exp[A_c(t-s)]\Delta A_c(s)x(s)ds\\&+\int_0^t \exp[A_c(t-s)]\Delta A_{co}(s)x(s-\tau)ds\end{aligned} \tag{223}$$

Next, use the real, positive, diagonal transformation

$$x(t) = D^{-1}z(t) \tag{224}$$

and the property that

$$\exp(DA_cD^{-1}t) = D\exp(A_ct)D^{-1} \tag{225}$$

and

$$\exp(A_ct) = M\exp(\Lambda t)M^{-1} = \sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} \tag{226}$$

to obtain

$$\begin{aligned} z(t) = & D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0) \\ & + \int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_c(s) D^{-1} z(s) ds \\ & + \int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_{co}(s) D^{-1} z(s-\tau) ds \end{aligned} \tag{227}$$

where M is a modal matrix of A_c; λ_i is the i-th eigenvalue of A_c with v_i and w_i the corresponding right and left eigenvectors, respectively; Λ is a diagonal matrix with the λ_i on the diagonal; and $(\cdot)^*$ denotes complex conjugate transpose. Note that

$$\|z(t)\| \to 0 \text{ implies that } \|x(t)\| \to 0 \tag{228}$$

So it is sufficient to prove that $\|z(t)\| \to 0$.

Since z(t) is causal, the second integral in Eq.(227) should start from τ.

$$\begin{aligned} z(t) = & D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0) \\ & + \int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_c(s) D^{-1} z(s) ds \\ & + \int_\tau^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_{co}(s) D^{-1} z(s-\tau) ds \end{aligned} \tag{229}$$

In the second integral of Eq.(229) let $u=s-\tau$, i.e. $s=u+\tau$, $ds=du$, then

$$z(t)=D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0)$$

$$+\int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_c(s) D^{-1} z(s)ds$$

$$+\int_0^{t-\tau} D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-u-\tau)} \Delta A_{co}(u+\tau) D^{-1} z(u)du \qquad (230)$$

Note that u is a dummy variable of integration. Therefore, we may substitute s for u in Eq. (230).

$$z^+(t) \le \left[D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i t} D^{-1} z(0)\right]^+$$

$$+\left[\int_0^t D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_c(s) D^{-1} z(s)ds\right]^+$$

$$+\left[\int_0^{t-\tau} D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s-\tau)} \Delta A_{co}(s+\tau) D^{-1} z(s)ds\right]^+$$

$$\le D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i t} D^{-1} z^+(0)$$

$$+\int_0^t \left[D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s)} \Delta A_c(s) D^{-1} z(s)\right]^+ ds$$

$$+\int_0^{t-\tau} \left[D\sum_{i=1}^{n} v_i w_i^* e^{\lambda_i (t-s-\tau)} \Delta A_{co}(s+\tau) D^{-1} z(s)\right]^+ ds$$

$$\le D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i t} D^{-1} z^+(0)$$

$$+\int_0^t D\sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s)} A_{cmax} D^{-1} z^+(s)ds$$

$$+ \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s-\tau)} A_{comax} D^{-1} z^+(s) ds \tag{231}$$

where $\alpha_i = -\mathrm{Re}[\lambda_i (A+BFC)]$, (232)

$$A_{cmax} = A_{max} + B_{max}(FC)^+ \tag{233}$$

$$A_{comax} = A_o^+ + A_{omax}(t) \tag{234}$$

and where we have used the property that $[\exp(\lambda_i t)]^+ = \exp(-\alpha_i t)$. Next, integrate both sides of Eq.(231) to obtain

$$\begin{aligned}
\int_0^\infty z^+(t)dt \le\ & D\sum_{i=1}^{n} (v_i w_i^*)^+ \int_0^\infty e^{-\alpha_i t} dt \cdot D^{-1} z^+(0) \\
& + \int_0^\infty \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s)} A_{cmax} D^{-1} z^+(s) ds dt \\
& + \int_0^\infty \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s-\tau)} A_{comax} D^{-1} z^+(s) ds dt \\
= & D\sum_{i=1}^{n} (v_i w_i^*)^+ \int_0^\infty e^{-\alpha_i t} dt \cdot D^{-1} z^+(0) \\
& + \lim_{R\to\infty} \left[\int_0^R \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s)} A_{cmax} D^{-1} z^+(s) ds dt \right] \\
& + \lim_{R\to\infty} \left[\int_0^R \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s-\tau)} A_{comax} D^{-1} z^+(s) ds dt \right] ; 0 \le s \le t
\end{aligned} \tag{235}$$

The order of integration in Eq.(235) may be interchanged because all of the functions inside the integrals are continuous functions of t and s Churchill (1972). Thus, Eq.(235) is equal to

$$\lim_{R\to\infty} \left[\int_0^R \int_0^t D \sum_{i=1}^{n} (v_i w_i^*)^+ e^{-\alpha_i (t-s)} A_{cmax} D^{-1} z^+(s) ds dt \right]$$

$$+ \lim_{R\to\infty}\left[\int_0^R\int_0^t D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i(t-s-\tau)} A_{comax} D^{-1} z^+(s)dsdt\right] \quad ; \quad 0\le s\le t$$

$$= \lim_{R\to\infty}\left[\int_{s=0}^t\int_{t=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i(t-s)} A_{cmax} D^{-1} z^+(s)dtds\right]$$

$$+ \lim_{R\to\infty}\left[\int_{s=0}^t\int_{t=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i(t-s-\tau)} A_{comax} D^{-1} z^+(s)dtds\right]; 0\le s\le t$$

(236)

Now use the change of variables given by $\gamma=t-\tau$, $d\gamma=dt$, $\gamma>0$. Then, Eq.(236) is equal to

$$\lim_{R\to\infty}\left[\int_{s=0}^t\int_{\gamma=-s}^{R-s} D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax} D^{-1} z^+(s)d\gamma ds\right]$$

$$+ \lim_{R\to\infty}\left[\int_{s=0}^t\int_{\gamma=-s}^{R-s} D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} e^{\alpha_i\tau} A_{comax} D^{-1} z^+(s)d\gamma ds\right]; 0\le s\le t, \gamma>0$$

$$\le \lim_{R\to\infty}\left[\int_{s=0}^t\int_{\gamma=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax} D^{-1} z^+(s)d\gamma ds\right]$$

$$+ \lim_{R\to\infty}\left[\int_{s=0}^t\int_{\gamma=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} e^{\alpha_i\tau} A_{comax} D^{-1} z^+(s)d\gamma ds\right]; 0\le s\le t, \gamma>0$$

$$= \lim_{R\to\infty}\left[\int_{\gamma=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} A_{cmax} D^{-1} d\gamma \int_{s=0}^t z^+(s)ds\right]$$

$$+ \lim_{R\to\infty}\left[\int_{\gamma=0}^R D\sum_{i=1}^n (v_i w_i^*)^+ e^{-\alpha_i\gamma} e^{\alpha_i\tau} A_{comax} D^{-1} d\gamma \int_{s=0}^t z^+(s)ds\right]; \ \alpha_i>0$$

$$\le \lim_{R\to\infty}\left[D\sum_{i=1}^n \frac{(v_i w_i^*)^+}{-\alpha_i} (e^{-\alpha_i R}-1) A_{cmax} D^{-1}\int_0^\infty z^+(s)ds\right]$$

$$+ \lim_{R\to\infty}\left[D\sum_{i=1}^n \frac{(v_i w_i^*)^+}{-\alpha_i} (e^{-\alpha_i R}-1) e^{\alpha_i\tau} A_{comax} D^{-1}\int_0^\infty z^+(s)ds\right]; \ \alpha_i>0 \qquad (237)$$

Evaluate the limit of the term outside the integral in Eq.(237), note that s is now a dummy variable of integration, recall that the α_i's are positive, and substitute the result into Eq.(235) to obtain

$$\int_0^\infty z^+(t)dt \le D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i} D^{-1}z^+(0)$$

$$+ D\left(\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax} + \sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i} e^{\alpha_i \tau} A_{comax}\right)D^{-1}\int_0^\infty z^+(t)dt \tag{238}$$

Take norms in Eq.(238) and rearrange to obtain

$$\left\|\int_0^\infty z^+(t)dt\right\| \le \frac{\left\|D\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i} D^{-1}z^+(0)\right\|}{1 - \left\|D\left(\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax} + \sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i} e^{\alpha_i \tau} A_{comax}\right)D^{-1}\right\|} \tag{239}$$

Thus, $\left\|\int_0^\infty z^+(t)dt\right\| < \infty$ if

$$\left\|D\left(\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax} + \sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i} e^{\alpha_i \tau} A_{comax}\right)D^{-1}\right\| < 1 \tag{240}$$

Note that

$$\left\|\int_0^\infty z^+(t)dt\right\| < \infty \text{ implies that } \int_0^\infty \|z^+(t)\|dt < \infty \tag{241}$$

Also note that x(t) and z(t) are uniformly continuous on $(0,\infty)$ because of the linearity of the uncertain closed loop plant which together with

Eqs.(240)-(241) implies that $\|z^+(t)\| \to 0$ as $t \to 0$ (Hsu and Meyer (19??). This implies that $\|x(t)\| \to 0$ as $t \to 0$ which proves that the linear uncertain closed loop plant is asymptotically stable.

Finally, Perron weightings may be used for the matrix D in Eq.(240) in the same manner as shown by Sobel et. al. (1989) for an earlier robustness result. Thus, to reduce conservatism, Eq.(240) may be replaced by

$$\lambda_{max}\left\{\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}A_{cmax}+\sum_{i=1}^{n}\frac{(v_i w_i^*)^+}{\alpha_i}e^{\alpha_i\tau}A_{comax}\right\} < 1 \qquad (242)$$

where $\lambda_{max}(\cdot)$ of a non-negative matrix denotes the real non-negative eigenvalue $\lambda_{max} \geq 0$ such that $\lambda_{max} \geq |\lambda_i|$ for all eigenvalues λ_i.

Q.E.D.

We remark that the term $\exp(\alpha\tau)$ appears in Eq.(240). Therefore, it is more difficult to satisfy the sufficient condition as the time delay τ becomes larger. If $\tau \to \infty$, the norm in Eq.(240) goes to infinity because α is positive for a stable system. This implies that the sufficient condition for robust stability cannot be satisfied for a system with infinite time delay.

5. CONCLUSIONS

A new sufficient condition for the robust stability of a linear time invariant system subject to time varying structured state space uncertainty has been proven. The proof is based upon showing that the integral of the norm of the uniformly continuous system state is bounded. A robust design method was proposed for the EMRAAT missile which minimizes the excursions in roll rate due to an initial sideslip subject to constraints on settling time, damping, control surface deflection rates, and robustness. The new robust design compares favorably with an earlier design which used eigenstructure assignment with an orthogonal projection for computing the achievable eigenvectors. In particular, the maximum of the absolute value of the integrated roll rate is reduced by approximately 80% when compared to the earlier design.

We have extended eigenstructure assignment to linear time invariant plants which are represented by the so-called unified delta model which is valid both for continuous time and sampled data operation of the plant. We have shown that the eigenvectors of the delta model are identical to the eigenvectors of the continuous time plant and an expression was derived for the eigenstructure assignment feedback gain matrix for the delta model. We have proposed a sufficient condition for the robust stability of a linear time invariant unified delta plant subject to linear time invariant structured state space uncertainty. A robust sampled data design is computed for the Extended Medium Range Air to Air Missile by minimizing the integral of the roll rate with

constraints on selected eigenvalues, actuator deflection rates, and the sufficient condition for robust stability. The robust design is compared with an orthogonal projection eigenstructure assignment design. In particular, the maximum integrated roll rate to a "1-cosine" sideslip gust is reduced from 18.87 degrees for the orthogonal projection design to 0.6990 degrees for the robust design.

Finally, we have presented two new sufficient conditions for the robust stability of a continuous time system with time delay which is subjected to linear time-varying structured state space uncertainty.

6. REFERENCES

Andry, A. N., Jr., Shapiro, E. Y., and Chung, J. C. (1983). Eigenstructure assignment for linear systems. *IEEE Transactions on Aerospace and Electronic Systems*, **AES-19**, 711-729.

Bossi, J.A. and Langehough, M.A. (1988). Multivariable autopilot designs for a bank-to-turn missile. *Proc. ACC*, Atlanta, GA, U.S.A., 567-572.

Churchill, R.V. (1972). *Operational Mathematics*. Third Edition,McGraw-Hill, New York.

Golub, G.H. and Van Loan, C.F., (1983). *Matrix Computations*. The Johns Hopkins University Press, Baltimore.

Grace, A.,(1990). *Optimization Toolbox for use with MATLAB™*. The Mathworks Inc., Natick, MA, U.S.A.

Hsu, J.C. and Meyer A.U. (1968). *Modern Control Principles and Applications*. McGraw-Hill, New York.

Kouvaritakis, B. and Latchman, H. (1985). International Journal of Control, **41**, 1381.

Langehough, M. A. and Simons, F. E. (1988). 6DOF simulation analysis for a digital bank to turn autopilot. *Proc. ACC*, Atlanta, GA, U.S.A., 573-578.

McRuer, D., Ashkenas, I., and Graham, D. (1973). *Aircraft Dynamics and Automatic Control*. Princeton University Press, Princeton,NJ.

Middleton, R.H. and Goodwin, G.C. (1990). *Digital Control and Estimation: A Unified Approach*. Prentice Hall Inc., Englewood Cliffs NJ.

Military Specification-Flying Qualities of Piloted Airplanes. MIL-F-8785C, ASD/ENESS, Wright-Patterson AFB,Ohio.

Ogata, K (1987). *Discrete Time Control Systems*. Prentice-Hall Inc., Englewood Cliffs NJ.

Sobel, K.M., Banda, S.S., and Yeh H.-H. (1989). Robust control for linear systems with structured state space uncertainty. *Int. J. Control*, **50**, 1991-2004.

Sobel, K.M. and Cloutier, J.R. (1991). Eigenstructure assignment for the extended medium range air to air missile. *Journal of Guidance, Control, and Dynamics*, Engineering Note, in press.

Yu, W. and Sobel, K.M., (1991). Robust eigenstructure assignment with

structured state space uncertainty. *Journal of Guidance, Control, and Dynamics*, **14**, 621-628.

Yu, W., Piou, J.E., and Sobel, K.M. (1991). Robust eigenstructure assignment for the extended medium range air to air missile. Proc. IEEE Conf. Decision and Control, Brighton, UK, 2976-2981.

EVALUATION OF MOS-CONTROLLED THYRISTOR (MCT)
AT 270 V DC
FOR STATIC AND DYNAMIC LOADS

December 1991

Prepared by

M.A. Choudhry (Principal Investigator)

and

M.Mubeen(Research Assistant)

School of Electrical and Computer Engineering

West Virginia University

Morgantown, WV 26505

Prepared for

Universal Energy Systems, Inc.

4401 Dayton-Xenia Road

Dayton, OH 45432

ACKNOWLEDGEMENTS

I wish to thank Joseph A Weimer and John Nairus of Aero Propulsion and Power Laboratory , Wright-Patterson AFB, Ohio for technical discussions and providing MOS-Controlled Thyristors and Driver Circuit as a loan for the completion of this project. The support provided by Harris Corporation in the form of MOS-Controlled Thyristors and financial support from Allegheny Power System is greatly appreciated. The help of Universal Energy Systems, Inc. in administering this contract is also acknowledged.

ABSTRACT

The voltage transients across the MOS-Controlled Thyristor(MCT) are reduced to less than 50 V at 270 V dc by proper selection of filtering capacitor and snubber circuit. The voltage/current characteristics of MCT are obtained with a shunt dc motor. The voltage/current characteristics of MCT with a dynamic load are different from the static load. The resistive load switched at a different frequency in parallel with an inductive load reduces the voltage transients. A digital simulation using PSpice is developed for static and dynamic loads. The digital simulation results verify the experimental results.

INTRODUCTION

Power electronic devices can be used for control and efficient utilization of electrical energy. However, these devices may be damaged by the electrical transients due to switching of the device or the load. Unlike Gate-Turn-off thyristor which can be turned on by applying a small positive pulse to the gate but requires about one fifth of normal current to turn-off, MOS-Controlled Thyristor (MCT) can be turned on and off with a very small amount of current pulse applied between gate and anode. This device combines the ease of MOS gate control with high voltage and high current capability of SCR [1]. MOS-Controlled thyristor has the potential to be used for medium and high power applications in "More Electric Aircraft" at 270 V dc. The voltage and current characteristics of MCT depend on the nature of the load, voltage, and the switching frequency [2].

Very large voltage transients appear with inductive load during turn-on of MCT without a snubber circuit. Voltage transients of less than 50 volts are observed with a proper choice of polarized snubber circuit and a filtering capacitor at 270 V dc. The voltage and current characteristics are also obtained with a shunt dc motor. A comparison is made between the digital simulation and the experimental results. The voltage waveforms for two static loads connected in parallel and switched at different frequencies are also obtained.

VOLTAGE AND CURRENT CHARACTERISTICS OF MCT WITH A STATIC LOAD

Figure 1 shows the single line diagram of the circuit used for evaluating the voltage/current characteristics of MCT at 270 V dc. A three-phase 15kVA, 240 V (line-to-line) ac generator supplies power to a three-phase diode-bridge rectifier. Two 800 μF capacitors are connected at the rectifier output to reduce the ripple in the dc voltage. A filtering capacitor and a polarized snubber circuit are connected at the MCT. A free-wheeling diode is connected across the load to reduce the switching voltage transients across the MCT.

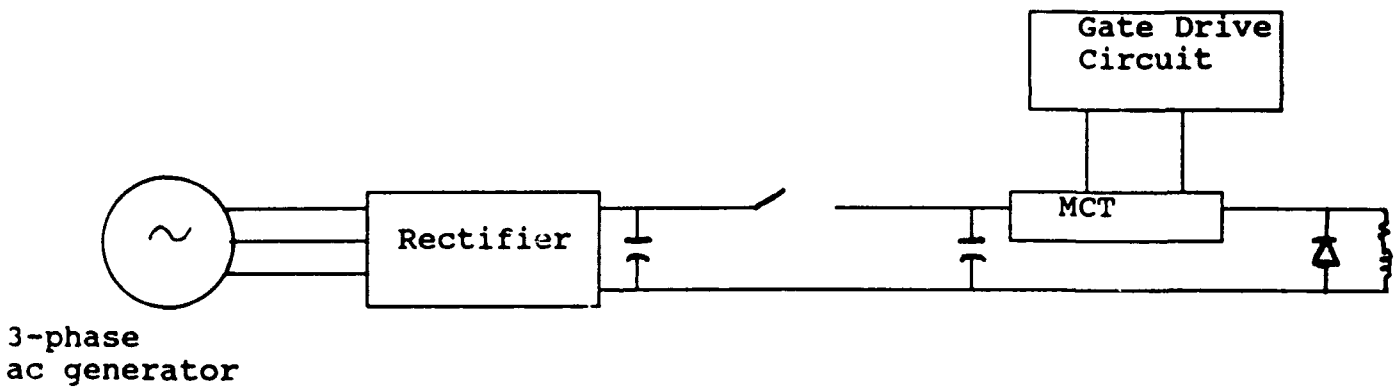


Fig. 1. Single line diagram of the circuit to evaluate voltage/current characteristics of MCT at 270 V dc with a static load.

Figure 2 shows the input and output voltage of gate drive unit. The input is a 0 to 5 V signal supplied from a function generator. Figure 3 shows a voltage waveform at a switching frequency of 1 kHz for a resistive load at 263 V dc. Figures 4 and 5 show the expanded voltage waveforms during turn-on and turn-off

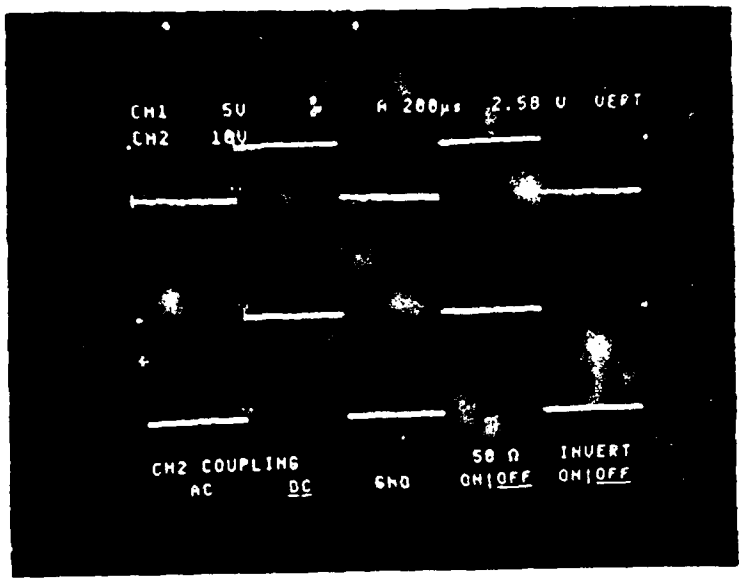


Fig. 2. Input and output voltage of gate drive circuit.
Top Trace: Input voltage 5V/cm, 0V=6cm
Bottom Trace: output voltage 10V/cm, 0V=3cm
Sweep: 200 µs/cm

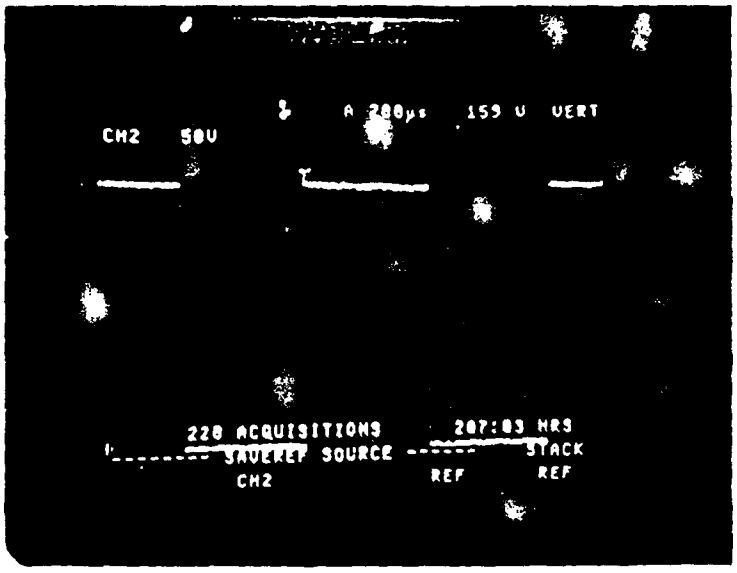


Fig. 3. Voltage waveform across the resistive load at a switching frequency of 1 kHz.
Vertical: 50V/cm, 0V=1cm Sweep: 200 µs/cm

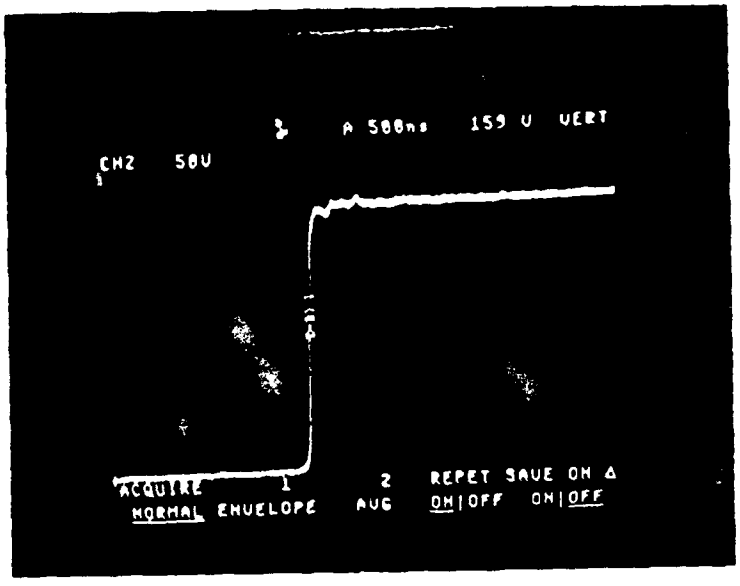


Fig. 4. Voltage waveform across the resistive load during turn-on of MCT.
Vertical: 50V/cm, 0V=1cm Sweep: 500 ns/cm

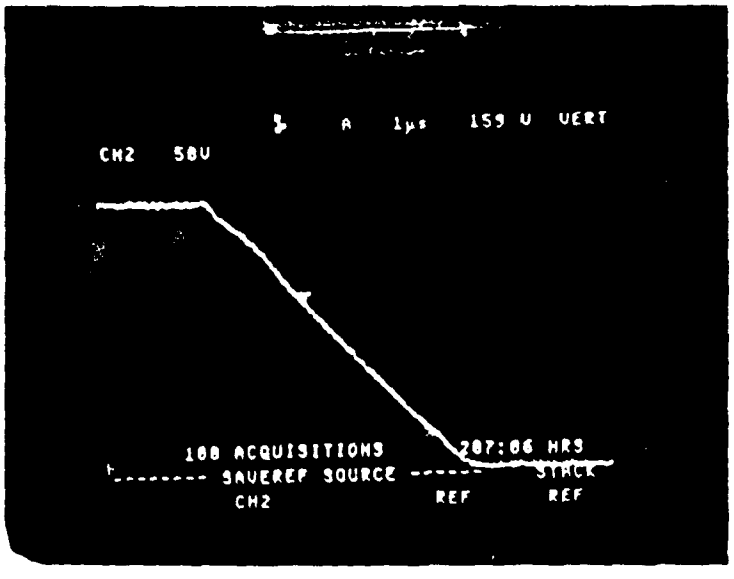


Fig. 5. Voltage waveform across the resistive load during turn-off of MCT.
Vertical: 50V/cm, 0V=1cm Sweep: 1 μs/cm

of MCT. Figure 6 shows a voltage waveform at a switching frequency of 1 kHz for an inductive load at 263 V dc. Figures 7 and 8 show the expanded voltage waveform during turn-on and turn-off of MCT with the inductive load. A small overshoot in voltage is observed during turn-on of MCT. However, no voltage transient is observed during turn-off of MCT.

Figure 9 shows the voltage and current waveforms in the inductive load at 260 V dc at switching frequency of 1 kHz. The current increases and decreases linearly during turn-on and turn-off of MCT due to the inductance in the load. Figure 10 shows the single line diagram of the circuit for digital simulation using PSpice.The supply voltage is modeled by an ideal source. Figure 11 shows the voltage waveform across the inductive load obtained from digital simulation at a switching frequency of 1 kHz. Figure 12 shows the expanded waveform during turn-on of MCT. Voltage transients only appear during turn-on of MCT. The digital simulation does not match exactly with the experimental results due to the difficulty of modeling the physical circuit elements in the digital simulation.

EFFECT OF FILTERING CAPACITOR ON TURN-ON TRANSIENT VOLTAGE OF MCT

The value of filtering capacitor of MCT has considerable effect on the voltage transients during turn-on of MCT with an inductive load. Figure 13 shows the voltage at the cathode of MCT during turn-on with C= 100.2 μF and 0.2 μF. A large transient voltage appears with C=0.2 μF. Figure 14 shows the cathode voltage of MCT during turn-on for C=200 μF and 100 μF. An increase in voltage overshoot is observed with C=200 μF. A C=100.2 μF appears the best

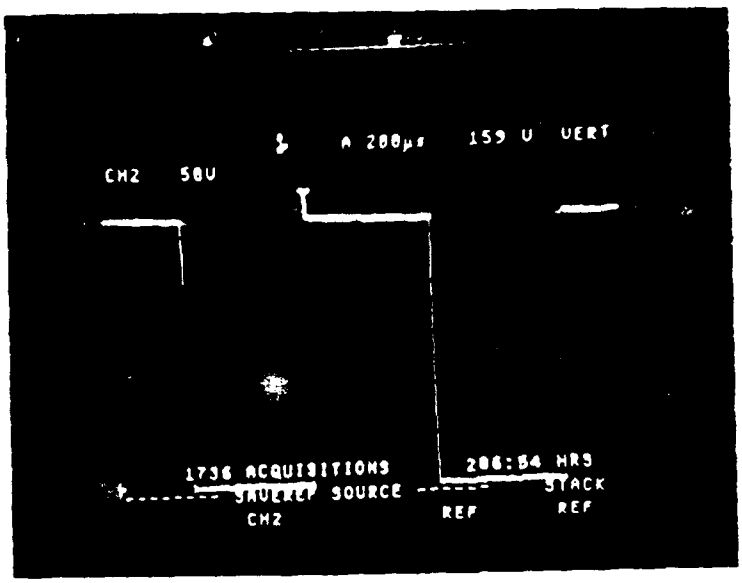


Fig. 6. Voltage waveform across the inductive load at a switching frequency of 1 kHz.
Vertical: 50V/cm, 0V=1cm Sweep: 200 µs/cm

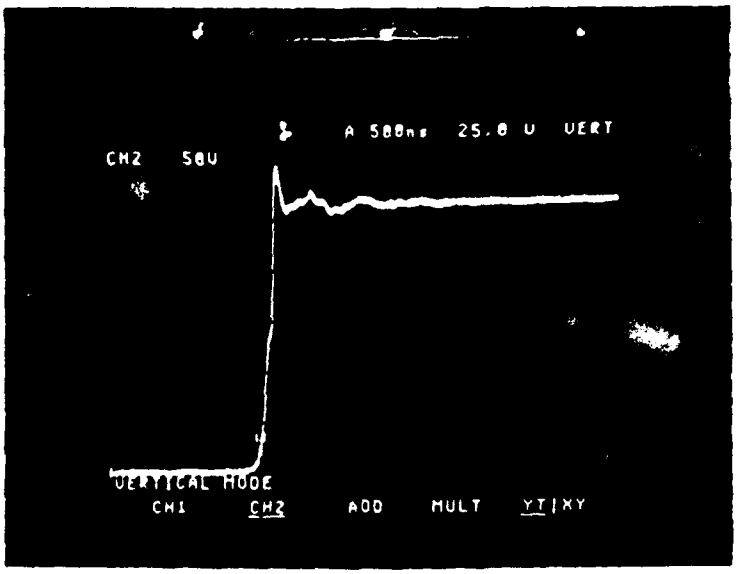


Fig. 7. Voltage waveform across the inductive load during turn-on of MCT.
Vertical: 50V/cm, 0V=1cm Sweep: 500 ns/cm

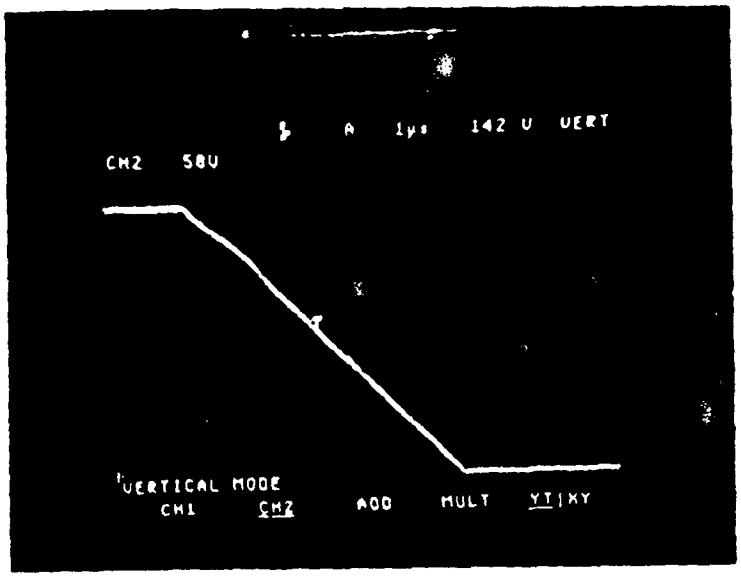


Fig. 8. Voltage waveform across the inductive load during turn-off of MCT.

Vertical: 50V/cm, 0V=1cm Sweep: 1 μs/cm

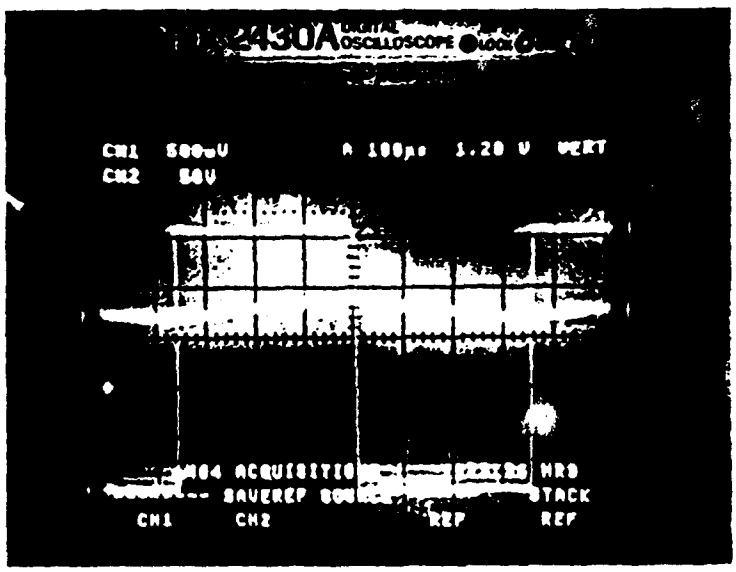


Fig. 9. Voltage and current waveforms across the inductive load at a switching frequency of 1.4 kHz.

Top Trace: Voltage
50V/cm, 0V=1cm
Sweep:100 μs/cm

Bottom Trace: Current
3.03A/cm, 0A=3cm
Sweep: 100 μs/cm

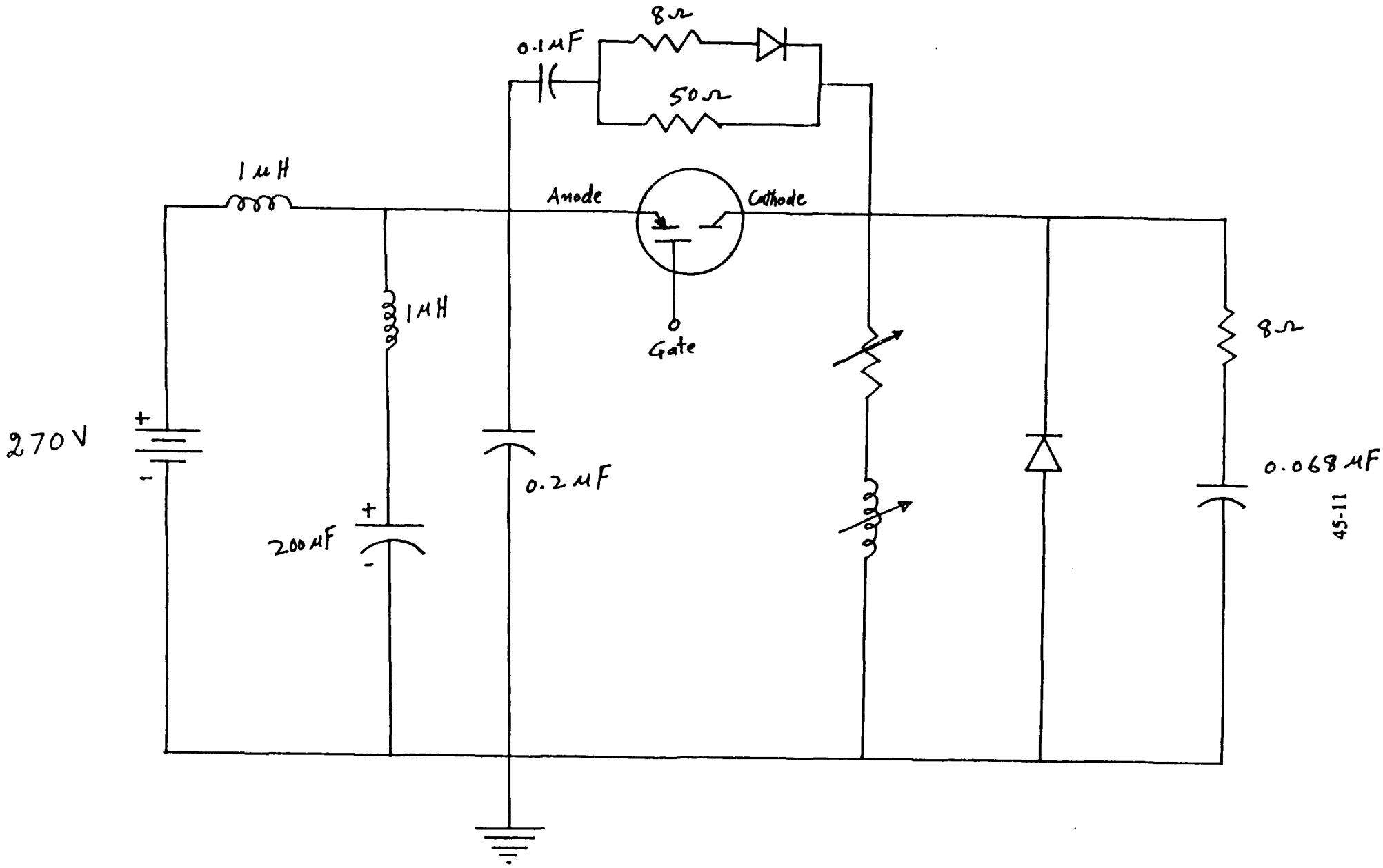


Fig. 10. Single line diagram of circuit for digital simulation.

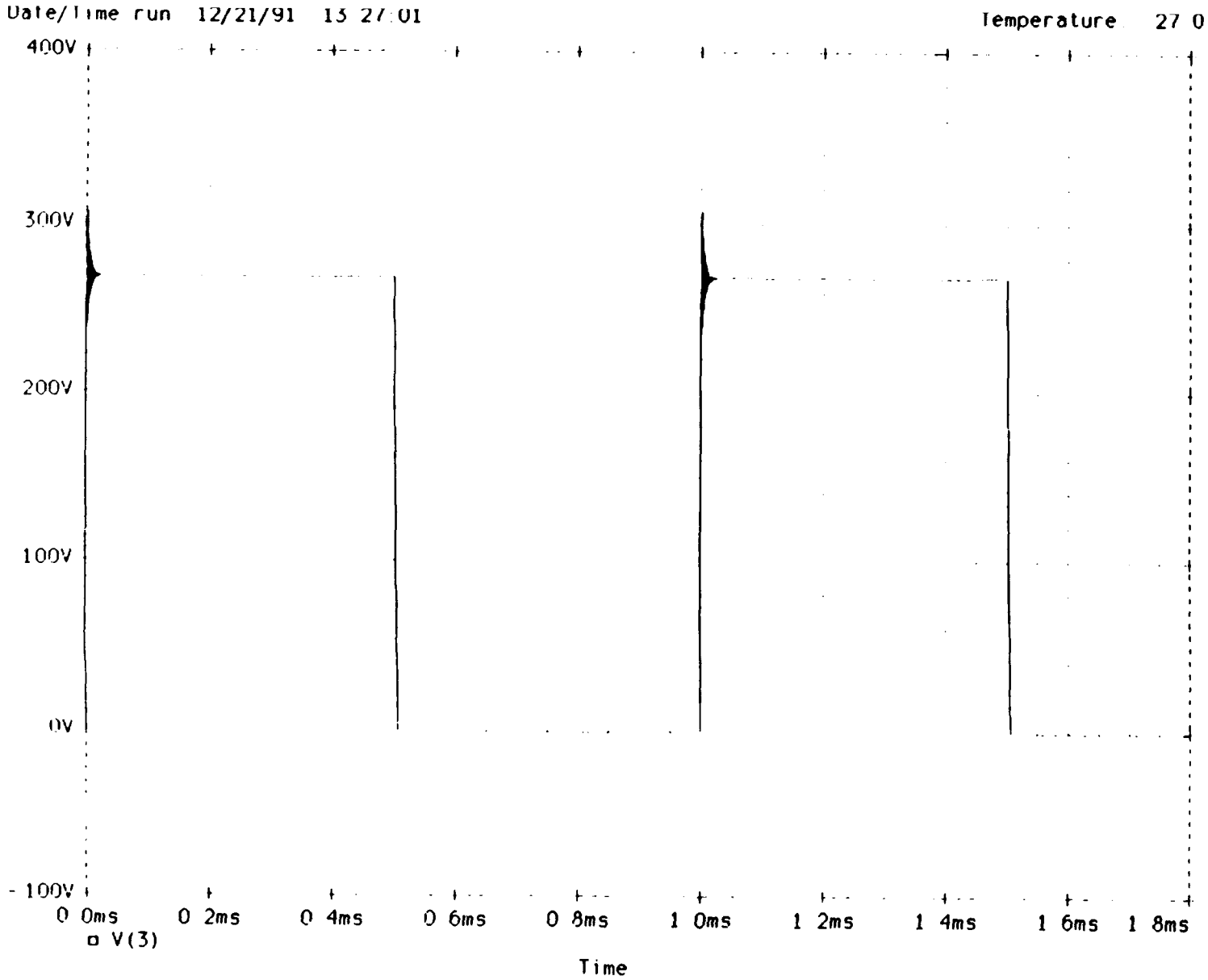


Fig. 11. Voltage waveform from digital simulation across the inductive load at a switching frequency of 1 kHz.

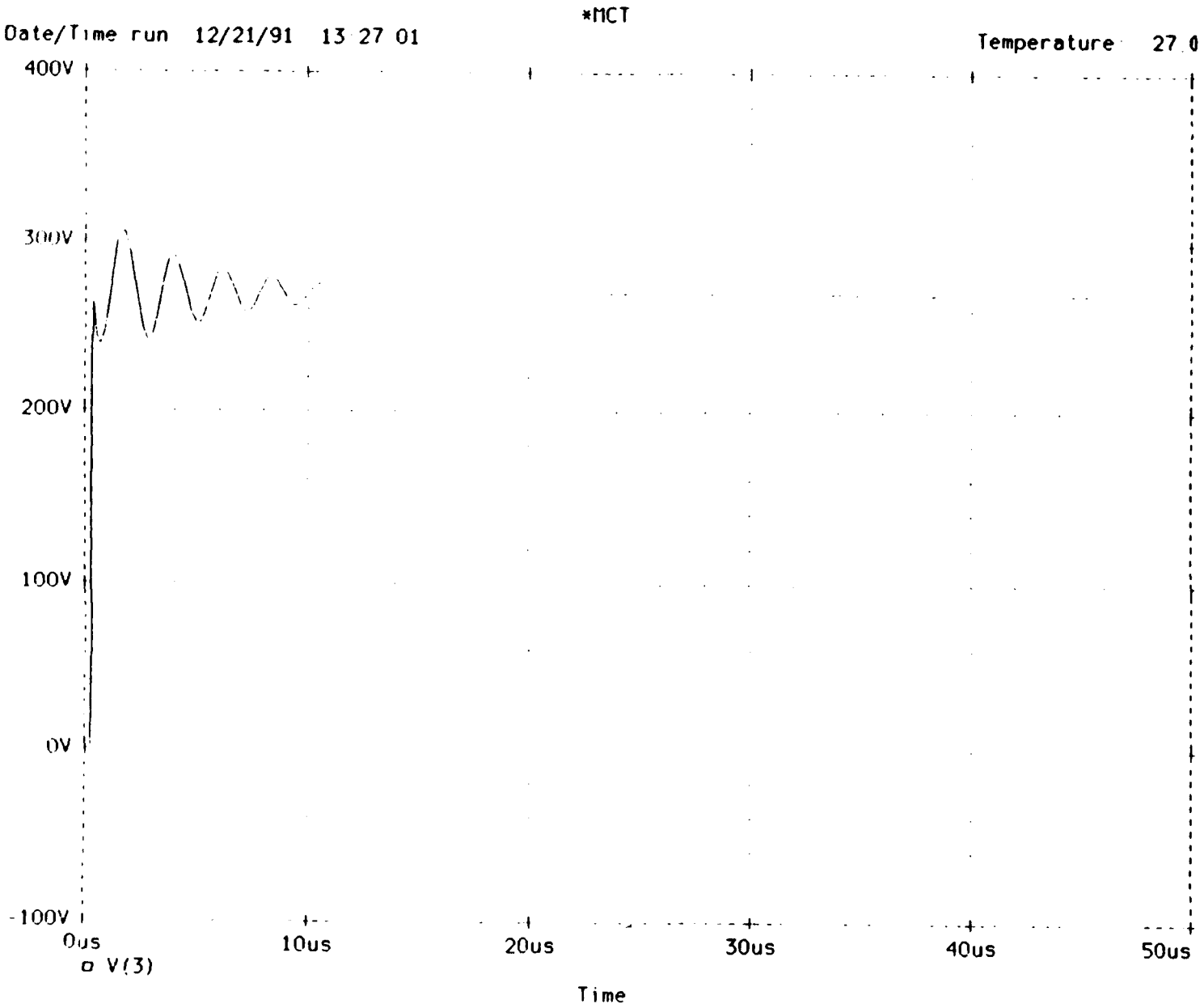


Fig. 12. Voltage waveform from digital simulation across the inductive load during turn-on of MCT.

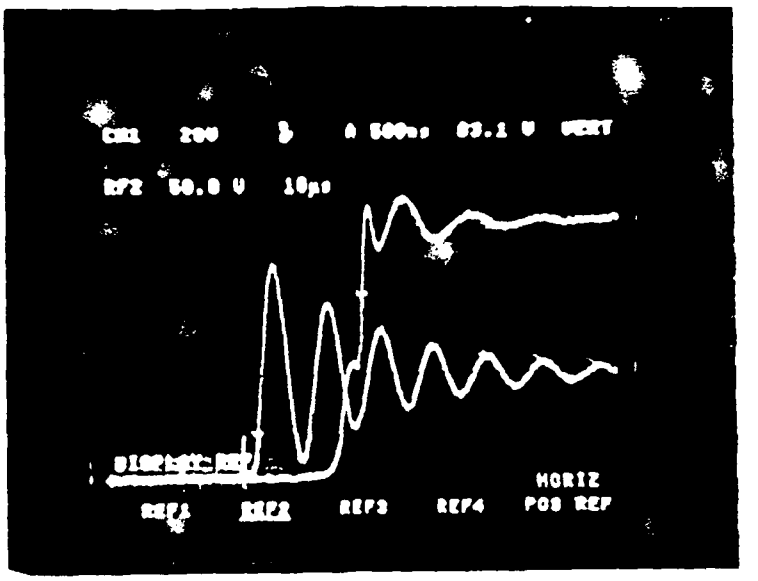


Fig. 13. Voltage waveforms across the inductive load at a switching frequency of 1 kHz.
Top Trace: 100.2 μF
20V/cm, 0V=1cm
Sweep:500 ns/cm
Bottom Trace: 0.2 μF
50V/cm, 0V=1cm
Sweep: 10 μs/cm

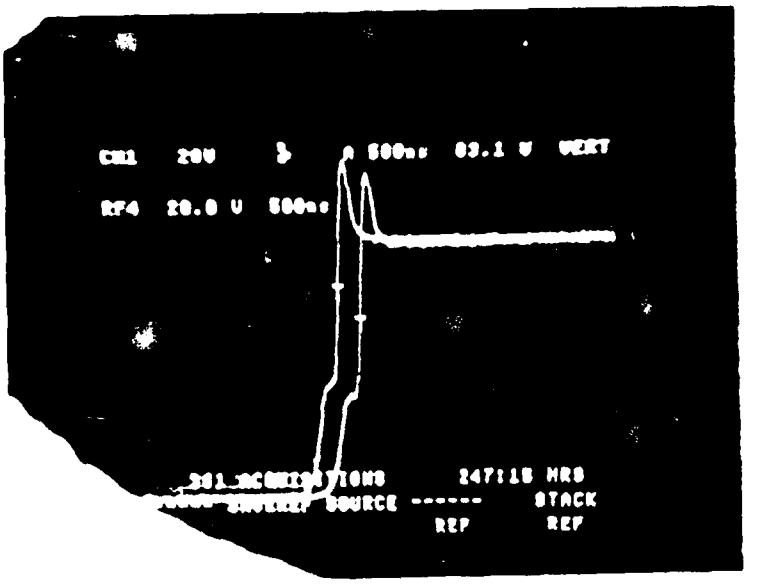


Fig. 14. Voltage waveforms across the inductive load at a switching frequency of 1 kHz.
Left Trace: 200 μF
20V/cm, 0V=1cm
Sweep:500 ns/cm
Right Trace: 100.2 μF
20V/cm, 0V=1cm
Sweep: 500 ns/cm

choice for the filtering capacitor. However, the digital simulation showed less overshoot in voltage with a higher value of filtering capacitor.

VOLTAGE AND CURRENT CHARACTERISTICS OF MCT WITH A DYNAMIC LOAD

Figure 15 shows the single line diagram of the circuit for evaluating the voltage/current characteristics of MCT with a dynamic load. A 7.5 hp, 240 V shunt dc motor is supplied current through a MCT. The MCT is turned on during the starting of the dc motor. Figure 16 shows the gate voltage of MCT and the voltage across the shunt motor. A voltage of 180 V is applied to the dc motor when the MCT is on. The terminal voltage of the motor goes to zero when the MCT is turned off. The terminal voltage of the motor remains zero for approximately 450 ns before increasing to 150 V. The terminal voltage goes to 180 V when the MCT is turned on again. Figure 17 shows the line current and the motor terminal voltage during MCT turn-on and turn-off. The line current immediately goes to zero when the MCT is turned off and increases linearly due to the armature inductance when the MCT is turned on. Figure 18 shows the armature current and the terminal voltage of the motor during MCT turn-on and turn-off. The armature current flows through the free-wheeling diode and decreases linearly when the MCT is turned off. The field current remains approximately constant during MCT turn-off period due to the large inductance of the field winding. Figure 19 shows the single line diagram of circuit for digital simulation using PSpice. The back emf of the motor is represented by a constant voltage source. Figure 20 shows the terminal voltage of the motor from the digital simulation when

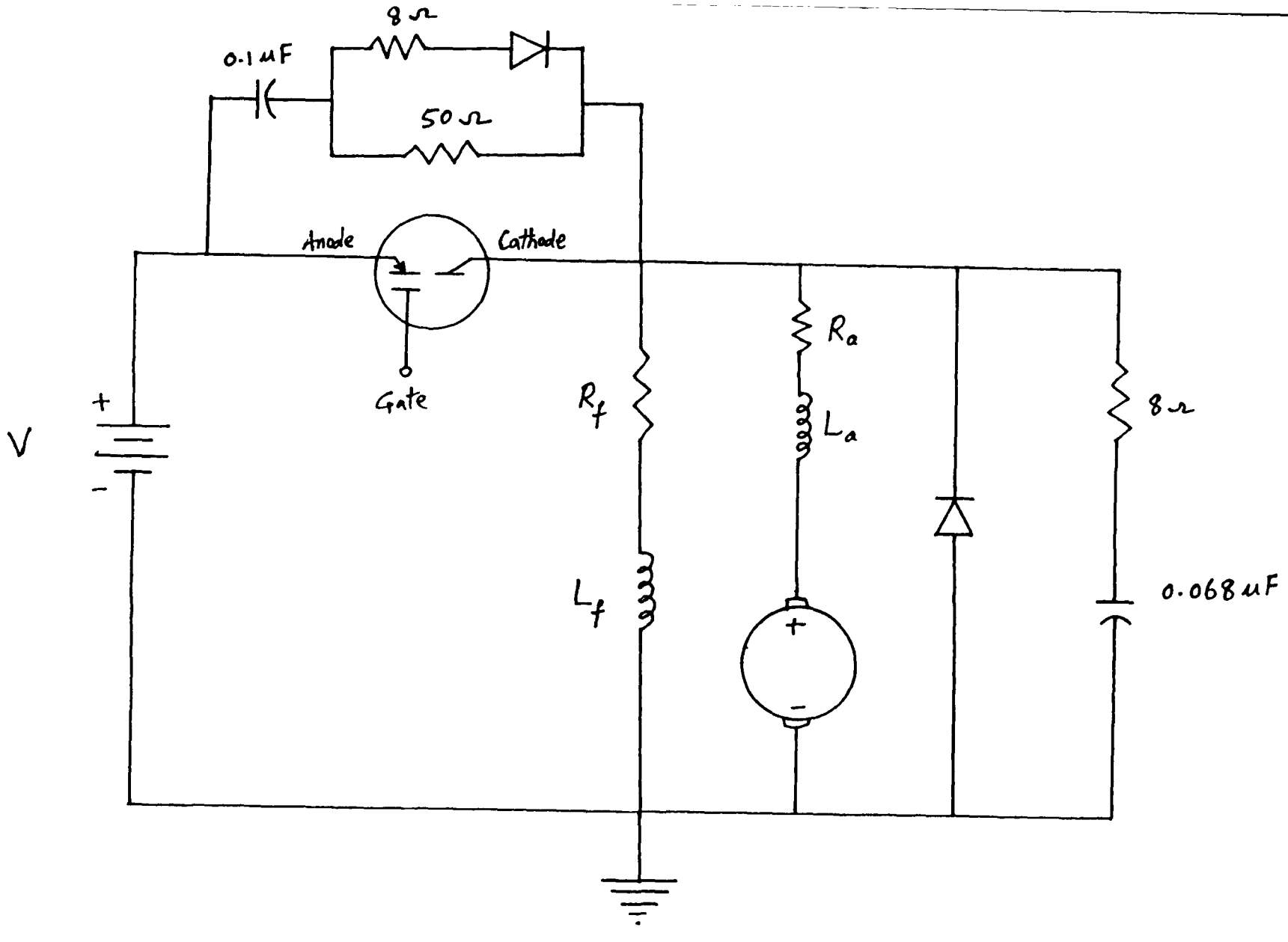


Fig. 15. Single line diagram of the circuit to evaluate voltage/current characteristics of MCT with a dynamic load.

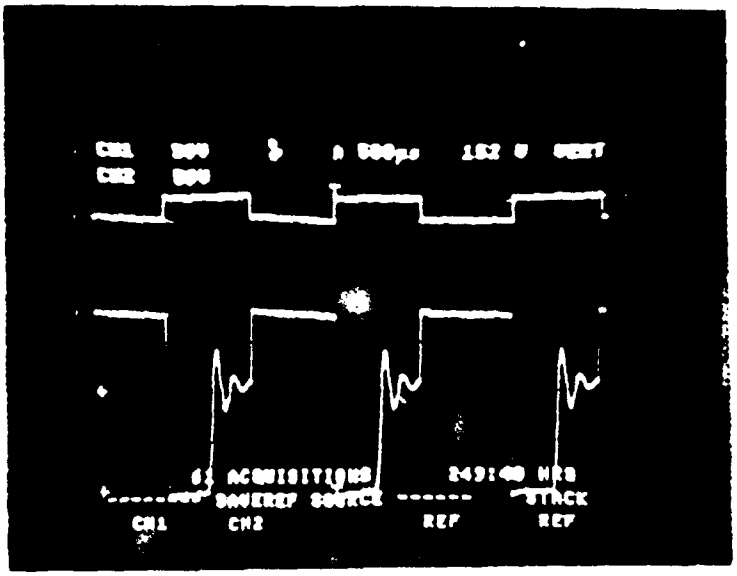


Fig. 16. Gate voltage of MCT and terminal voltage across the motor.

Top Trace: Gate voltage	Bottom Trace: Voltage
50V/cm, 0V=3cm	50V/cm, 0V=1cm
Sweep:500 μs/cm	Sweep: 500 μs/cm

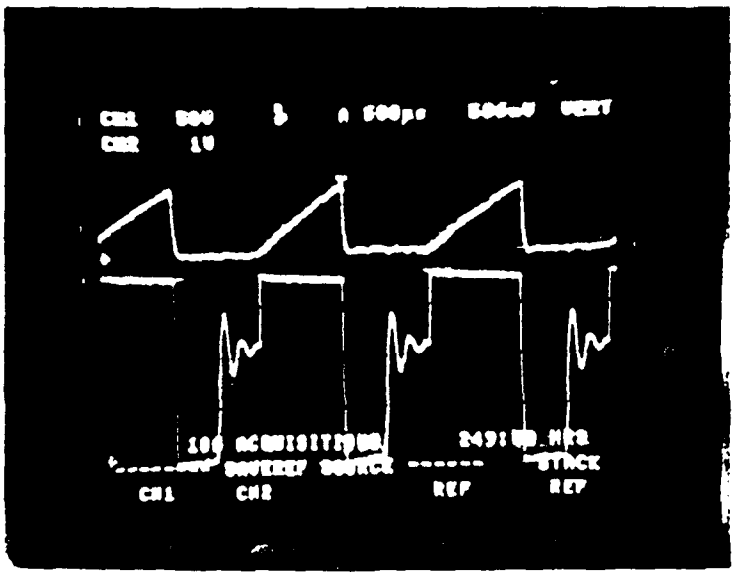


Fig. 17. Line current and terminal voltage across the motor.

Top Trace: Current	Bottom Trace: Voltage
6.06A/cm, 0A=5cm	50V/cm, 0V=1cm
Sweep:500 μs/cm	Sweep: 500 μs/cm

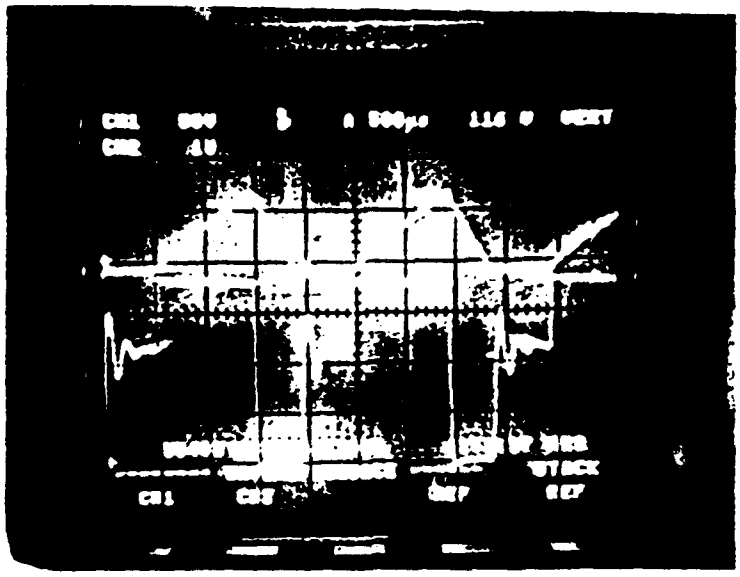


Fig. 18. Armature current and terminal voltage across the motor.

Top Trace: Current	Bottom Trace: Voltage
6.06A/cm, 0A=5cm	50V/cm, 0V=1cm
Sweep:500 µs/cm	Sweep: 500 µs/cm

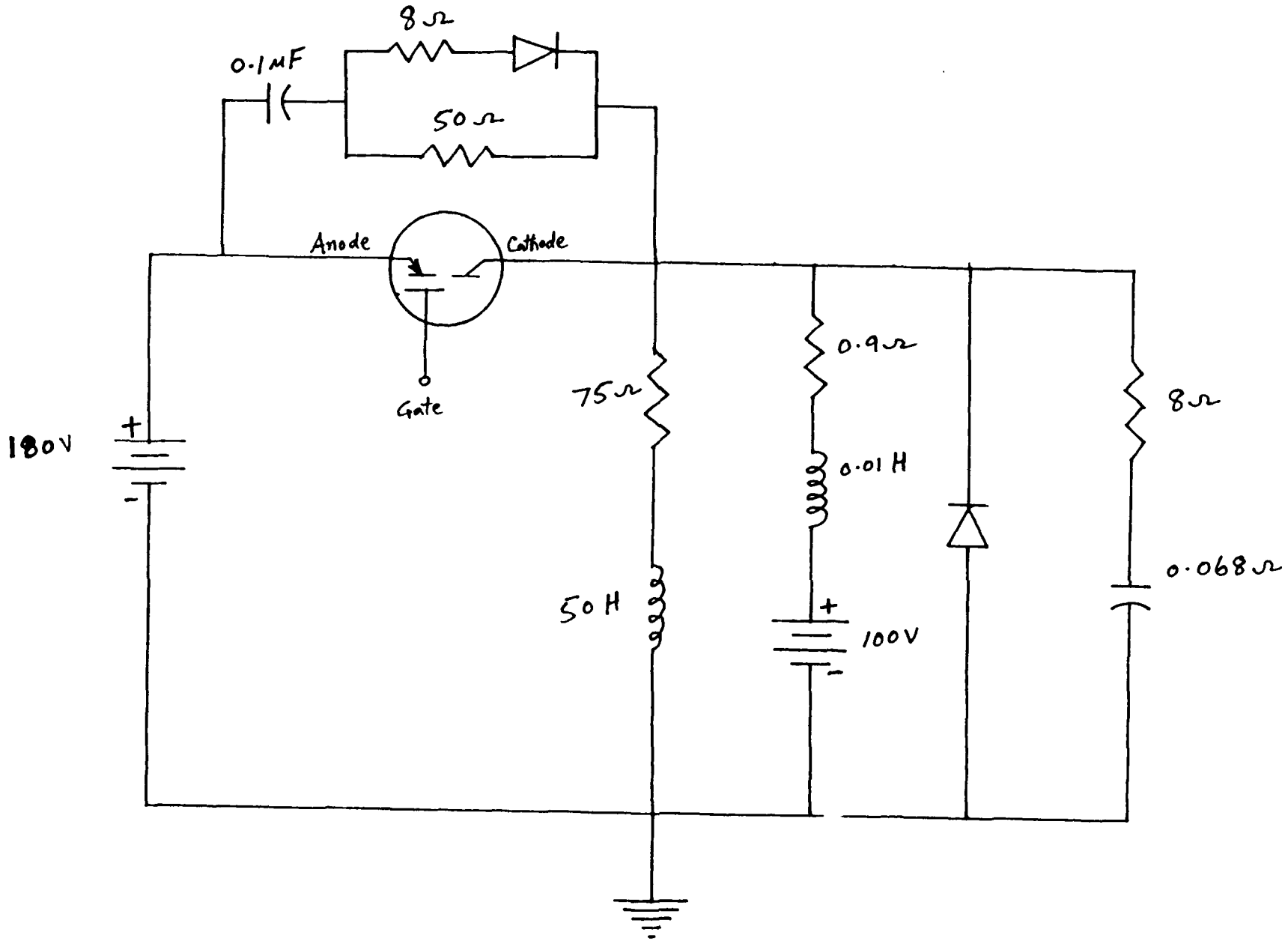


Fig. 19. Single line diagram of the circuit for digital simulation of dynamic load.

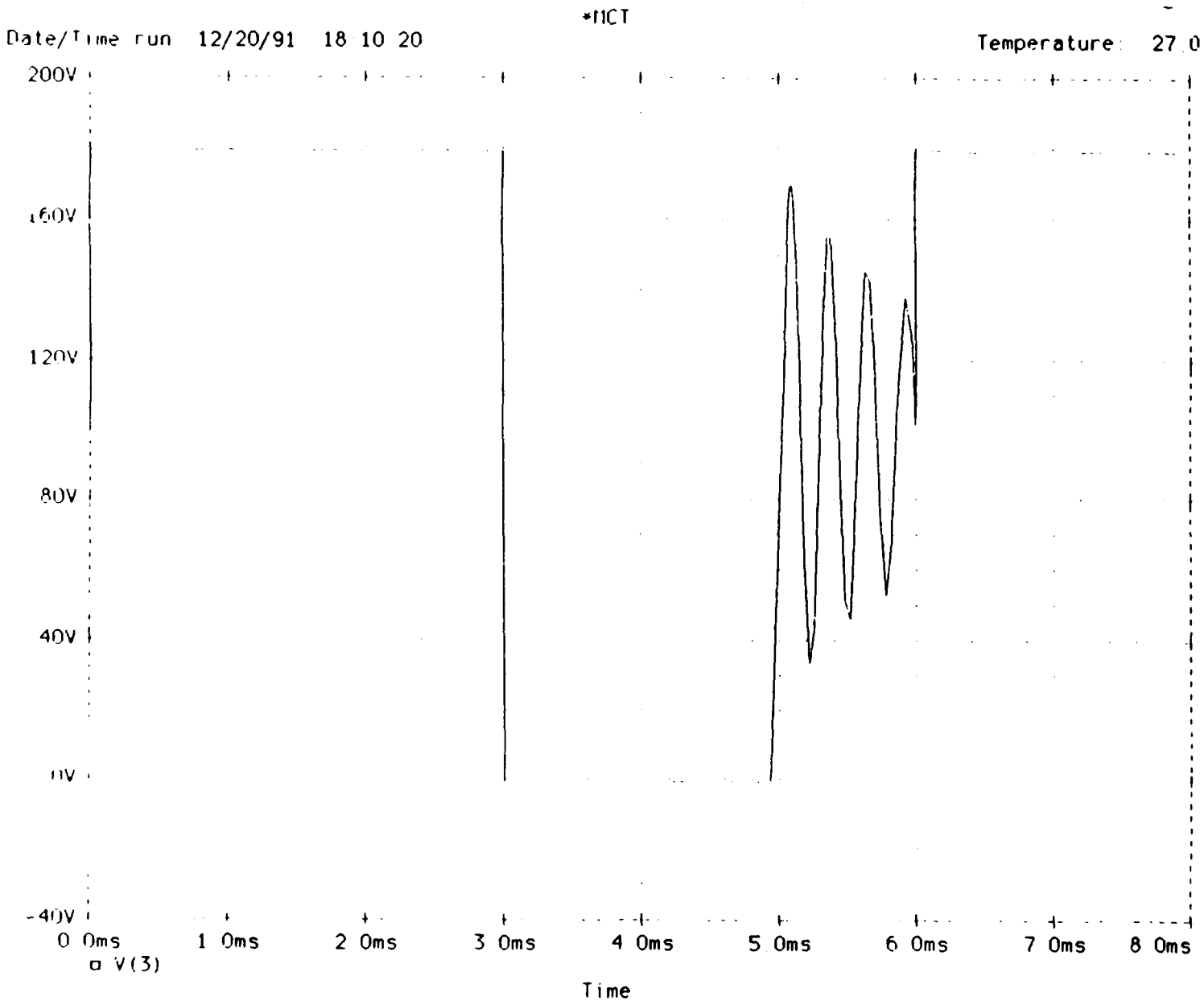


Fig. 20. Terminal voltage of motor from digital simulation.

the MCT is switched at 167 Hz. The voltage waveform from digital simulation is similar to the experimental results.

PARALLEL LOAD SWITCHING AT DIFFERENT FREQUENCIES

Figure 21 shows the voltage waveforms across the resistive and the inductive loads connected in parallel. The resistive load is turned on and off by the first MCT at a frequency of 1 kHz and the inductive load is switched at 1.5 kHz by the second MCT. Figure 22 shows the voltages across the two load during turn-on. The turn-on time of the MCT supplying the inductive load is greater than the MCT supplying the resistive load. The addition of the resistive load in parallel to the inductive load reduces the turn-on transient voltage of the inductive load.

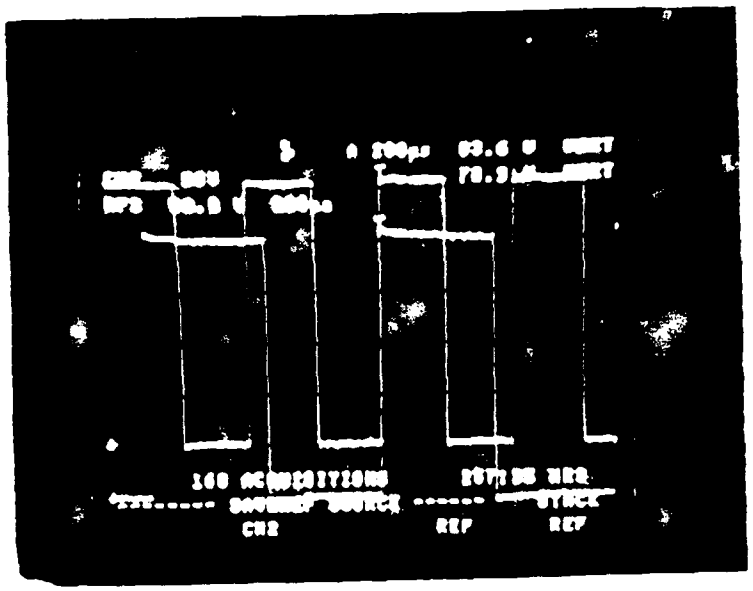


Fig. 21. Voltage waveforms across the resistive and inductive loads connected in parallel.
Top Trace: Inductive load
50V/cm, 0V=2cm
Sweep:200 μs/cm

Bottom Trace:Resistive load
50V/cm, 0V=1cm
Sweep: 200 μs/cm

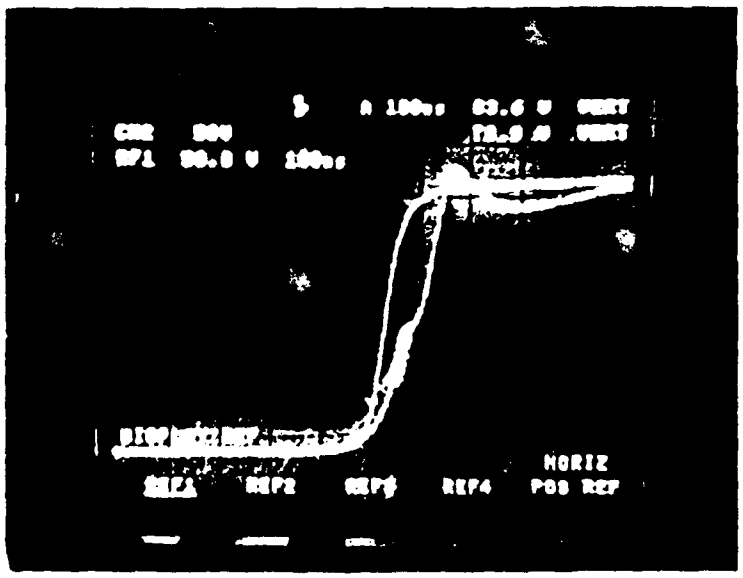


Fig. 22. Voltage waveforms across the resistive and inductive loads connected in parallel during turn-on.
Left Trace: Resistive load
50V/cm, 0V=1cm
Sweep:100 ns/cm

Right Trace:Inductive load
50V/cm, 0V=1cm
Sweep: 100 ns/cm

REFERENCES

[1]. V.A.K. Temple,"Advances in MOS-Controlled Technology," PCIM.,November 1987, pp. 12-15.

[2]. M.A.Choudhry,"Evaluation of MOS-Controlled Thyristor(MCT) at 270 Volt DC for Resistive and Inductive Loads," Energy Systems, Inc. Dayton, Ohio, August, 1990.

Final Report

Non Newtonian Effects of Powder Lubricant Slurries In Hydrostatic and Hydrodynamic Bearings

Funded by

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
BOLLING AFB, DC

Delivered to

Universal Energy Systems, Inc.
ATTN: Mr. Rodney C. Darrah
4401 Dayton-Xenia Road
Dayton, OH 45432

by

Dr. Don W. Dareing
Professor, Mechanical Engineering
University of Florida
Gainesville, FL 32611

December 31, 1991

ACKNOWLEDGEMENT

This research was supported by the Air Force Office of Scientific Research, Bolling AFB, DC. The work is a direct extension of studies made by the principle investigator during the summer of 1990 under a summer faculty research program. The funding of this project expanded tribology research in the Mechanical Engineering Department at the University of Florida and supported two Master of Science graduate students. We are most grateful to the Air Force Office of Scientific Research for funding this research.

Non Newtonian Effects of Powder Lubricant Slurries In Hydrostatic and Hydrodynamic Bearings

by
Dr. Don W. Dareing

OBJECTIVE OF THE RESEARCH

The objective of the research was to evaluate the behavior of powder lubricant slurries in thick film lubrication applications. The powder lubricant was graphite and the carrier fluid was ethylene glycol. The evaluation was based on laboratory testing and mathematical formulations accounting for non Newtonian rheological properties of these types of slurries.

Laboratory test rigs were designed and fabricated to evaluate the slurry in a hydrodynamic journal film, a hydrostatic fluid film and a squeeze film. Fluid film pressure measurements were also taken for ethylene glycol alone to establish a baseline set of data for the evaluation. Even though the potential application of this particular powder lubricant slurry is for high temperature environments the testing during this pilot program was conducted at room temperatures.

One of the goals was to formulate fluid film pressure equations which would account for the non Newtonian behavior of powder lubricant slurries. Mathematical pressure formulations were developed concurrently with the collection of laboratory pressure data.

SIGNIFICANCE OF THE RESEARCH

The Department of Defense, NASA, and the US gas turbine industry have recently embarked on a major new initiative, the Integrated High Performance Turbine Engine Technology (IHPTET) program. The goal of IHPTET is to double propulsion capability shortly after the turn of the century. With the advent of IHPTET has come the requirement for significantly advanced gas turbine engine lubricants and lubrication system components.

One approach currently being pursued in trying to meet the IHPTET requirements is the use of powder solid lubricants for the various lubrication system components. These powders could be supplied to the components via a pressurized air stream or through pump pressurized slurries containing the poser lubricant.

In developing the powder lubricant slurry approach, it is clear that the fundamental lubrication principles must be fully understood. The proposed research will provide fundamental insight into the rheological effects of powder slurries on hydrostatic, squeeze films and hydrodynamic lubrication films.

The results from the research will provide a basis for formulating the more complex problem of elastohydrodynamic (EHD) lubrication of rolling contact bearings and gears.

METHOD OF ANALYSIS

Much of the effort during this one year project was directed at the design and fabrication of laboratory test rigs. These rigs were needed to experimentally determine the effects of powder lubricants on the lubrication fluid film pressure of a carrier fluid. The carrier fluid used for this study is ethylene glycol.

Two test rigs were built. One test rig is based on a hydrodynamic journal bearing. The other test rig is designed to develop hydrostatic and squeeze films. These test rigs were designed and fabricated during the first half of 1991. They were debugged and instrumented during the third quarter of the year. The mathematical analysis, data collection and report writing took place during the last quarter of 1991.

Two graduate students, Mr. Zhenming Wu and Mr. Anup Batra, were assigned to work on this project under the direction of the principle investigator, Dr. D. W. Dareing. Even though the funding of this research by the Air Force will formally end on December 31, 1991, these two students will continue to collect and analyze data during the spring semester of 1992 and use the results of this research as the basis for their Masters Thesis.

DESCRIPTION OF LABORATORY TEST EQUIPMENT

Laboratory testing was an important aspect of this one year project because it helped to give insight into the behavior of powder lubricant slurries in thick film lubrication applications. Laboratory testing also gave both faculty and students an opportunity to observe lubrication film performance and to expand experimental skills in this area of research.

Two test rigs were designed and built especially for this research project. One rig was designed to study lubricants in a hydrostatic condition. This rig was also used to observe film pressures in a squeeze film mode. The second rig was designed to study lubricants under a hydrodynamic condition. This rig is basically an instrumented laboratory journal bearing.

These two laboratory test rigs are described below.

Hydrostatic Test Rig

The hydrostatic test rig (shown in Figures A1 and A2) contains two parallel surfaces separated by a lubricant. The lubricant is supplied to the surfaces by an external pump through an internal recessed pocket. The pump has the following specifications:

Make	- Micropump Corporation
Model	- 415
Motor/Pump Coupling	- Magnetic Drive
Maximum Flow Rate	- 630 ml/min
Rated Maximum Pressure	- 20 psi

The active film is dictated by the geometry of the upper surface which has an outside diameter of 4 inches and an inside diameter of 2 inches. The thickness of the film is constant for a given test but this film thickness can be changed by means of a fine adjustment thread (48 NF). This fine thread connects the upper bearing pad to the frame.

The two surfaces in this particular test rig do not rotate so that the temperature within the fluid film does not change appreciably from the bulk temperature of the lubricant. The main reason for fixing the two surfaces was limited funds for the project. This feature of the test rig, however, allowed precise measurements of predetermined settings in film thicknesses by using fixed LVDT displacement sensors;

make - Trans-Tek, Inc.
model - DC-DC Series 240
Model 02141-0000

The upper thrust bearing pad is mounted within its vertically moveable support so that adjustments can be made to the plane of the bearing surface. Adjustment screws are mounted at each of three LVDT locations to maintain parallelism between the upper and lower bearing surfaces. This feature was very important in collecting consistent data because parallelism of the surfaces greatly affected volume flow rate between the bearing surfaces.

Laboratory equipment in the Mechanical Engineering Department at the University of Florida does not include a profilometer. Therefore it was not possible to measure surface waviness or surface roughness. However, pressure and flow rate measurements indicate some waviness in the surfaces as measured flow rates were higher than calculated flow rates for film thickness settings less than 0.002 inches.

Fluid film pressures are measured at four (4) different location across the film. These pressure points are located

a) within the recess pocket
b) 1/4 inch from the inside pocket
c) 1/2 inch from the inside pocket
d) 3/4 inch from the inside pocket

The pressure at the outside radius of the film is ambient. Fluid pressures at these points are measured by dial pressure gages.

Flow rates through the bearing is measured by use of a rotameter flow meter. The flow meter was calibrated for pure ethylene glycol and for the powder slurry. The calibration was made by pumping the liquid into a beaker of known volume and recording time.

Hydrodynamic Test Rig

The hydrodynamic test rig (shown in Figure A3 and A4) contains a rotating journal and a nonrotating bearing. The outside diameter of the journal is 2.0 inches. The radial clearance, c, is 0.003 inches and clearance ratio $c/r = 0.003$. The length of the active journal bearing film is three (3) inches. The nonrotating bearing is mounted on a vertically moveable support. The vertical position of the bearing relative to the journal can be adjusted by means of a fine thread (48 threads per inch). This feature allows the fluid film geometry to be changed and fixed for a given test. By moving the bearing relative to the journal, the minimum film thickness can be set and fixed for a given test.

The vertical position of the bearing relative to the shaft is measured by use of an LVDT displacement transducer. This LVDT is used to establish desired minimum film thicknesses between the journal and bearing.

The journal shaft is mounted on taper roller bearings that are fixed to the frame of the rig.

The shaft is driven by an electric motor:

make	- Dayton Electric Mfg Co.
model	- 2M168B
volts	- 90 volts DC
current	- 5.5 amps
speed	- 1725 rpm
power	- 0.5 hp

A variable speed control allows the motor speed to be varied continuously up to about 1725 rpm. Output motor is stepped up by use of a 4 to 1 timing gear pair. Journal shaft speed can therefore go as high as 6900 rpm with this drive arrangement.

Lubricant is supplied to the test journal bearing at the top side of the bearing by a low volume/low pressure pump. The lubricant drains into the pump from a funnel type reservoir located beneath the bearing.

Small holes are drilled through the bearing to provide access for measuring fluid film pressure around the journal. A diagram showing these pressure ports is given in Figure 11. These pressure ports are located in a transverse plane across the center of the bearing. In addition, one pressure port is located off of this center plane to indicate the effects of end leakage on fluid film pressure.

A thermocouple was placed in the bearing to measure the effects of fluid film friction on local temperature in the film. The temperature rise was significant and greatly affected the viscosity of the lubricants. This temperature rise is discussed in a later section of the report.

A special pressure transducer was made to measure fluid film pressures around the journal. The design of the pressure transducer was suggested by Professor A. Seireg. The pressure transducer consisted of a small cylinder containing strain gauges cemented to its surface. The small cylinder expanded as a thin walled cylinder when subjected to pressure. Surface strains could be calibrated to indicate pressure. This pressure transducer was inexpensive, easy to make and gave accurate pressure readings.

MATHEMATICAL FORMULATIONS

The mathematical formations in this section were developed to predict the pressure distributions in hydrostatic and squeeze films. Hydrostatic pressure equations were used to predict pressures in the hydrostatic laboratory test rig. A comparison of predicted and measured data are discussed in the next section. Squeeze film pressure equations and test data are not yet available for this report; it will be gathered this spring by Mr. Wu. In addition, hydrodynamic pressure equations have not been developed because of time constraints. In this case, Reynold's differential equation of lubrication is non linear. Mr. Batra will focus on the solution of this equation during the spring term.

Hydrostatic Film Equations

The hydrostatic pressure equations below are developed for the thrust bearing configuration shown in Figure 1. Because the different properties of Newtonian and non-Newtonian fluids, two different mathematical models are applied.

For ethylene glycol, Newtonian fluid model is applied. The following equation describes the pressures profile under beneath the working surface of the bearing.

$$p = \frac{6\mu Q}{\pi h_o^3} \ln \frac{R}{r} \qquad (1)$$

where

- p: Pressure (psi) at any point along the radius of the bearing,
- R: Radius (in) of the bearing,
- r: Any radius (in) between R_o and R,
- Q: Flow rate (in^3/sec),
- h_o: Film thickness (in),
- μ: Viscosity (reyns).

This equation is from Theory and Practice of Lubrication For Engineers, by **Dudley D. Fuller**, (see ref. 1)

For the mixture of graphite powder and ethylene glycol, Power Law Model is applied. The following equation shows relation between flowrate Q and the related parameters.

$$Q = 2b \left(\frac{1}{K}\frac{dp}{dx}\right)^{\frac{1}{n}} \left(\frac{h_o}{2}\right)^{\frac{2n+1}{n}} \left(\frac{n}{2n+1}\right) \qquad (2)$$

where

K: Power Law constant,
n: Power Law exponent.

This equation is given in "Non-Newtonian Behavior of Powder Lubricants Mixed with Ethylene Glycol", by **Dareing and Dayton**, (ref. 3). Replacing b by $2\pi r$, dx by dr, in solving for pressure gives

$$p = Q^{n} K \left(\frac{4\pi n}{2n+1}\right)^{-n} \left(\frac{h_o}{2}\right)^{-(2n+1)} \left(\frac{R^{1-n} - r^{1-n}}{1-n}\right) \qquad (3)$$

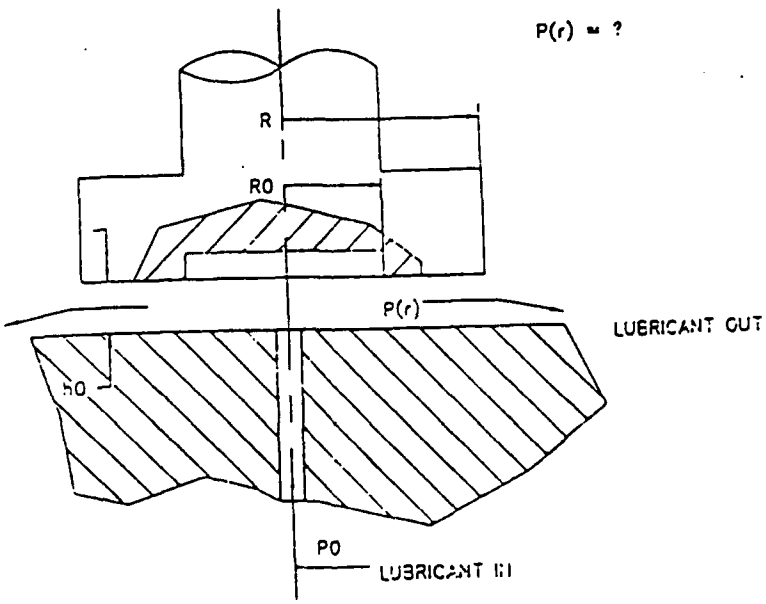


Figure 1 - Hydrostatic Thrust Bearing Geometry

PRESENTATION OF THE RESULTS

Initial testing with the two laboratory test rigs was conducted with pure ethylene glycol as the lubricant. The data from this initial testing provided a check on equipment and instrumentation through the comparison of measured pressure data with analytical prediction of pressure based on well established equations of lubrication for Newtonian fluids. The experimental data obtained by using pure ethylene glycol also provided a baseline from which to compare the effects of adding powdered graphite to the carrier fluid, ethylene glycol.

Laboratory data collected with both hydrostatic and hydrodynamic test rigs are presented below. The data includes measurements for pure ethylene glycol and for the graphite slurry.

Measurements of Hydrostatic Pressure and Flow Rates

Discussion of measurements with pure ethylene glycol.

Initial testing with pure ethylene glycol indicated the importance of surface geometry and parallelism between the two mating bearing surfaces. The effects of these factors on pressure and flow rate through the bearing were magnified when the minimum film thickness is of the order of 0.001 and 0.002 inches. Surface geometry was not measured because of lack of equipment. However parallelism of the two mating surfaces was controlled by three adjusting screws.

To control the parallelism of the mating surfaces, three LVDT's (linear variable differential transformer) were attached to the circular edge of one bearing 120 deg apart. Adjusting screws were also located at these points. Since the planar position of a surface is determined by three points, properly adjusting these devices allowed the film thickness to be controlled manually.

The readings of three LVDT will change whenever a change of film thickness happens. If this happens, by adjusting the adjustment devices until the readings of LVDT return to their original values, the film thickness is controlled.

However, because of the un-flatness of the surfaces of the bearing and the base of the rig, even though the position of the bearing surface can be controlled ideally, the surface waviness causes unevenness of the film thickness. This was a serious problem which affects the flow-rate and pressure relations.

With small film thickness, the effect of surface unflatness is apparent and the relationship between input pressure and flow-rate deviated significantly from theoretical predictions. But with larger film thickness such as 0.004 or 0.005 inch, the influence of surface unflatness on film pressure and flow rate is not so strong.

A rotameter was used in the test rig system to measure flow through the hydrostatic test bearing. This flow meter was calibrated simply by measuring volume flow into a beaker and recording time for this flow to take place. The calibration curve for pure ethylene glycol is given in Figure 2.

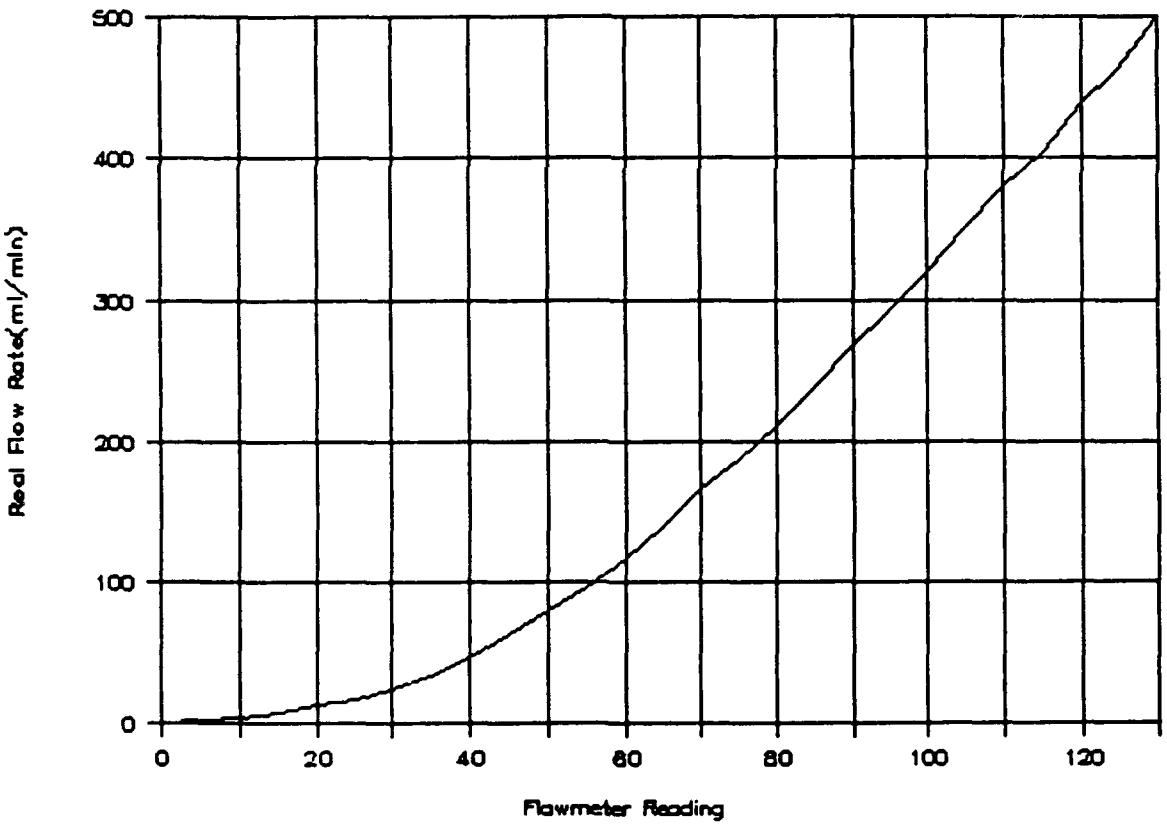


Figure 2 - Flow Meter Calibration Curve for Pure Ethylene Glycol

Since the flow-rate is less than 630 ml/min and both bearing surfaces are nonrotating, temperature increases due to fluid friction was negligible.

The purpose of the initial testing was to experimentally determine the relation between input pump pressure and flow rate between the bearing surfaces. This experimental data was then compared with theory predictions. The predicted input pressure vs flow-rate (Figure 3) is linear according to the equations given earlier, i.e., with the increase of flow-rate Q, the inlet pressure of the bearing p is proportionally increased.

The experimental data shown in Figure 3 closely match the predicted data. There is however a little deviation at high values of flow-rate and input pressure. This difference is possible due to the influence of surface unflatness. The obtained data show that a flow-rate of 200 ml/min requires a 14 psi input pressure if the film thickness is 0.004 inches. But if the film thickness is increased to 0.005 inches, a flow-rate of 211 ml/min requires only an input pressure of 7.5 psi.

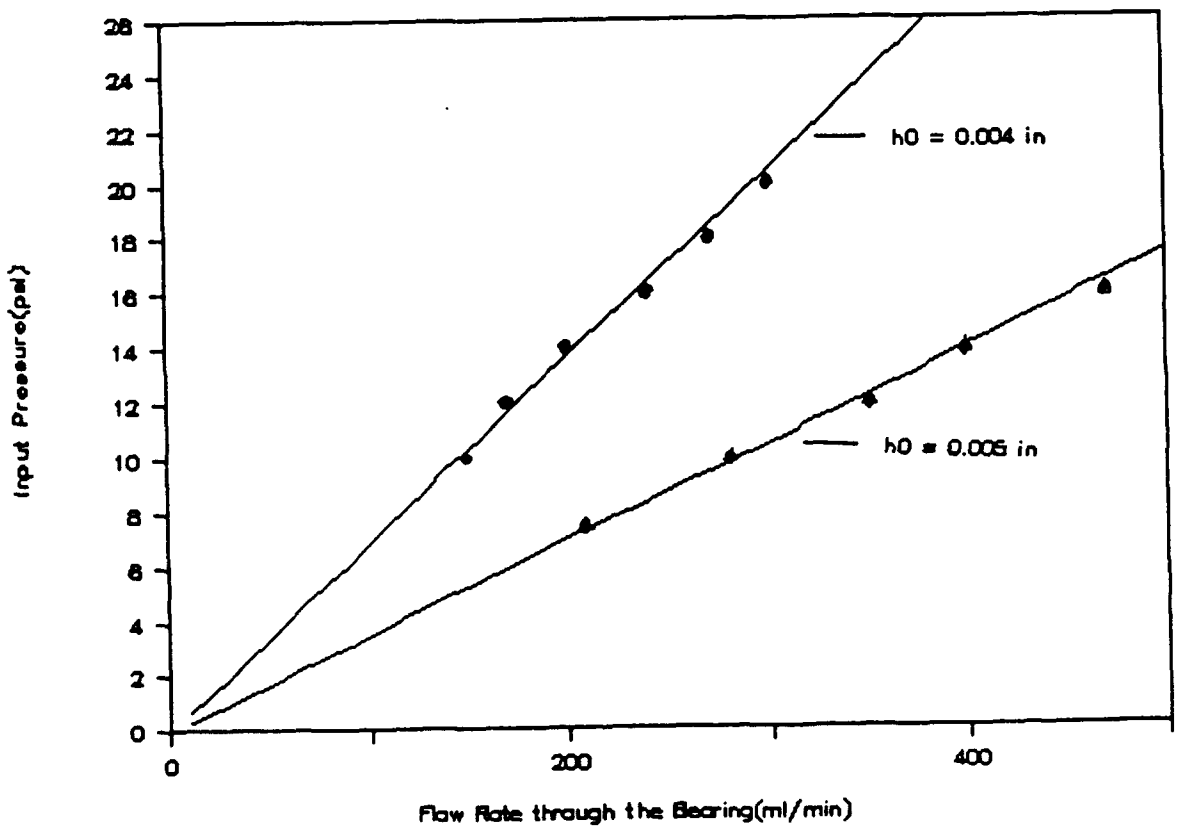


Figure 3 - Pump Pressure verses Flow Rate
(Pure Ethylene Glycol)

The predicted input pressure vs radius (Figures 4 and 5) is a logarithm curve. The theoretical curves are based on the equations given in the mathematical portion of the report. While the measured pressure in the recess closely match the predicted input pressure, measured pressures elsewhere are somewhat lower than predicted. This difference is probably due to the unevenness of the bearing surfaces.

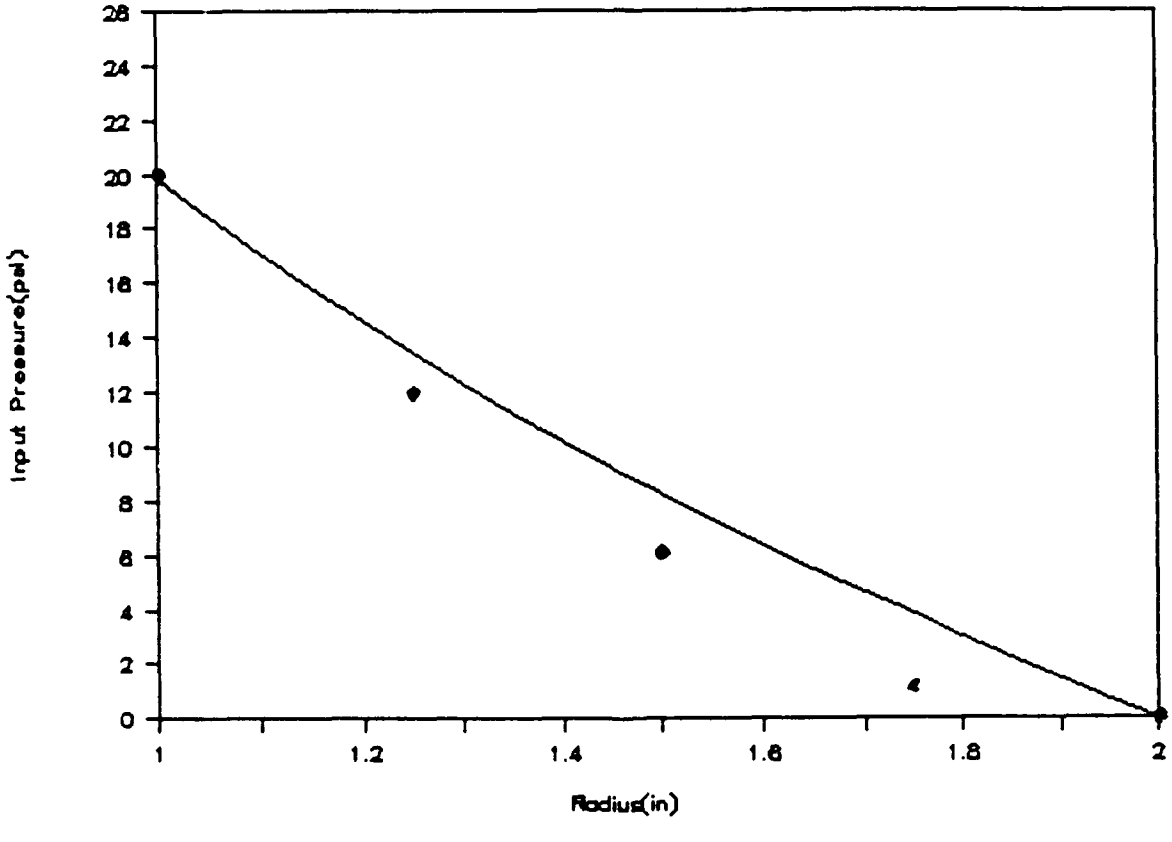


Figure 4 - Hydrostatic Pressure Distribution
- Pure Ethylene Glycol
- Film Thickness = 0.004 inches
- Flow Rate = 297 ml/min

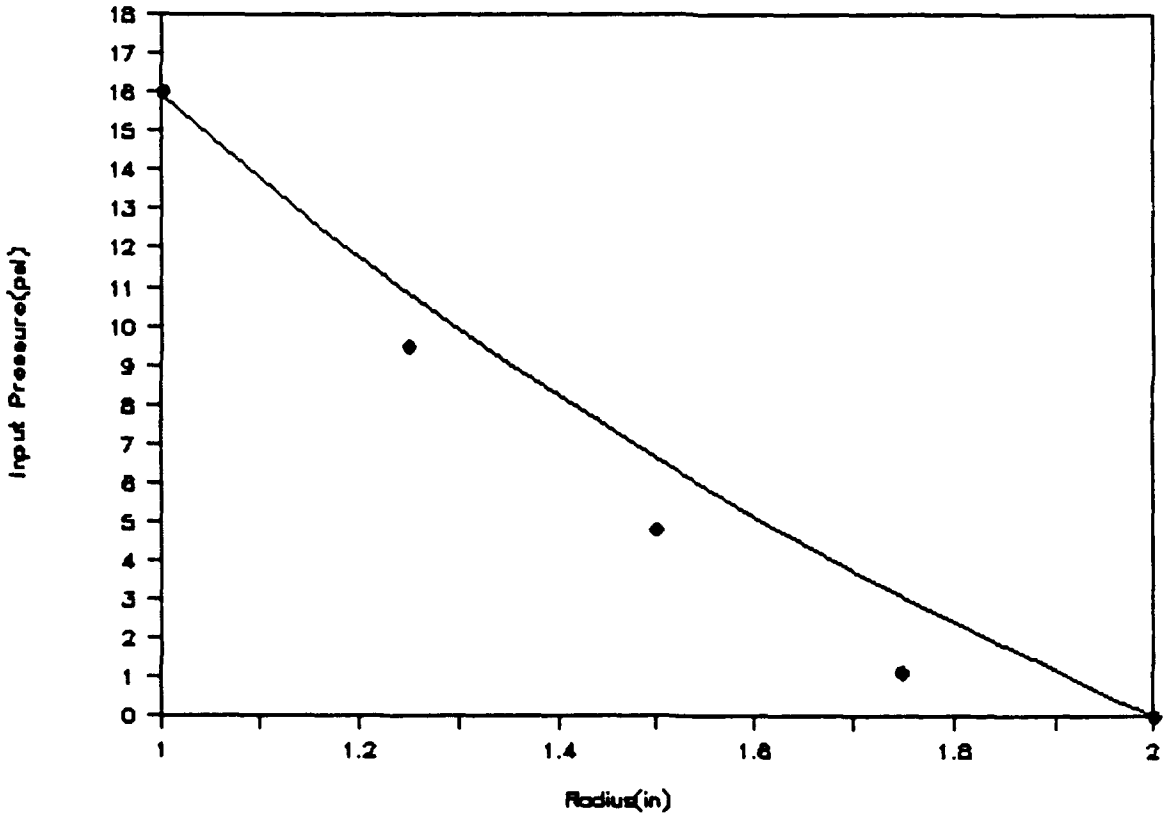


Figure 5 - Hydrostatic Pressure Distribution
* Pure Ethylene Glycol
* Film Thickness = 0.005 inches
* Flow Rate = 475 ml/min

Discussion of measurements with the graphite/ethylene glycol slurry

The slurry mixture is made of graphite powder and ethylene glycol having a weight ratio of one part by weight of graphite powder to eight part of ethylene glycol. The rheological properties of the slurry mixture are given by **Dareing and Dayton** (ref.3). Their work shows the non-Newtonian features of the slurry.

To minimize the effects of settling on the results, fresh slurry mixtures were used periodically.

The rotameter had to be recalibrated for the (1:8) slurry mixture. The calibration procedure in this case was the same as for the pure ethylene glycol. The calibration curve corresponding to this particular slurry mixture is given in Figure 6.

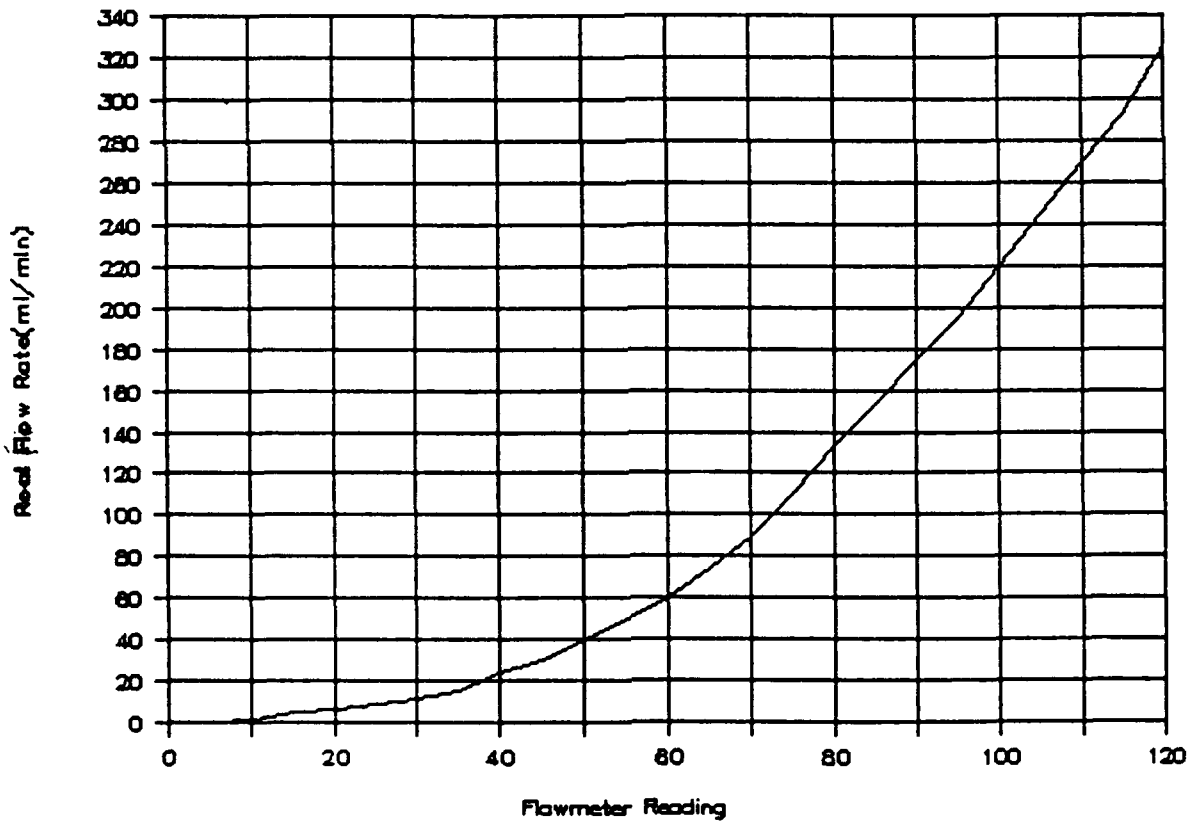


Figure 6 - Flow Meter Calibration Curve for Powder Slurry
(1:8 Graphite/Ethylene Glycol Mixture)

Film thickness h_o is the most decisive parameter in the system because it weighs nearly a third power of a given value (see the mathematical discussion section). Therefore, the flow-rate Q is very sensitive to any changes of the film thickness. A little increase or decrease of film thickness will cause large increase or decrease of flow-rates. This is proved by test data.

The relation between input pressure p and flow-rate Q is not linear for the powder lubricant slurry (Figure 7). This is one of the main differences between the performances of Newtonian and non-Newtonian fluids. If other parameters are fixed, the input pressure is proportional to the nth power of flow-rate Q. In this case, n is less than one.

The measured data (also in Figure 7) have a good match while the flow-rate is small, and have some deviation when the flow rate becomes larger (Q is about 140 ml/min when h_o is 0.004 inches; Q is about 240 ml/min when h_o is 0.005 inches).

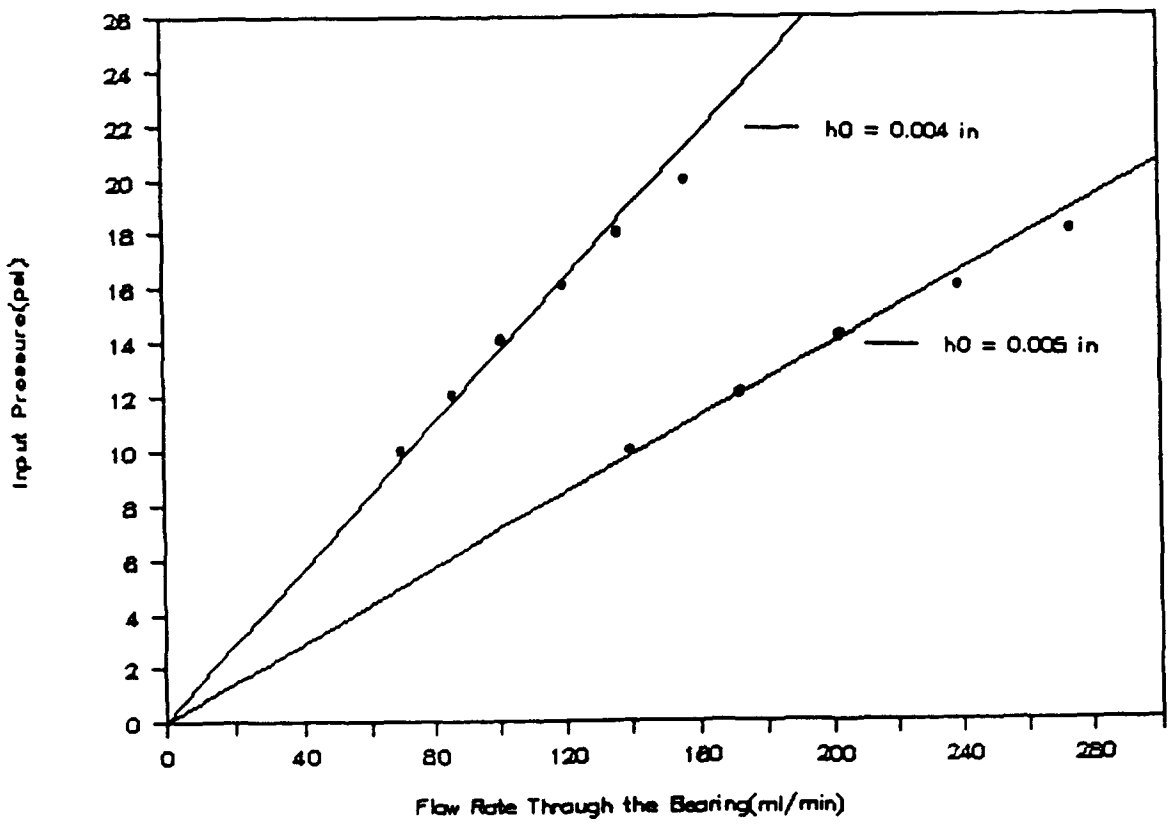


Figure 7 - Pump Pressure verses Flow Rate
(1:8 Graphite/Ethylene Glycol Mixture)

There are several reasons for these deviations. First, the changes of viscosity due to the settling of graphite powder particles, makes the loads carrying ability of the slurry lower than the theoretically calculated. Second, the unevenness of film thickness due to the unflatness of bearing surface. This is the main reason of the deviation because this parameter has a power factor nearly to three.

The predicted pressure profile along the radius of the bearing (Figures 9 and 10) is approximately proportional to a negative nth power of any value of r between the radian of the recess and the edge of the bearing, according to Power Law. The power factor n is less than one. The measured data (also in Figures 9 and 10) also show a close match to the predicted data. Though there are little deviations from the predicted curve, the tendency of pressure changing along with the radius is clear. The unevenness of film thickness contributes the deviation.

The data clearly show that the slurry mixture has the load carrying ability twice as much as for pure ethylene glycol. This means that the apparent viscosity of the graphite slurry mixture is about twice that of pure ethylene glycol. If the slurry mixture is used in a heavy duty hydrostatic thrust bearing system, the system will have a larger mechanical stiffness as a result of the addition of the graphite.

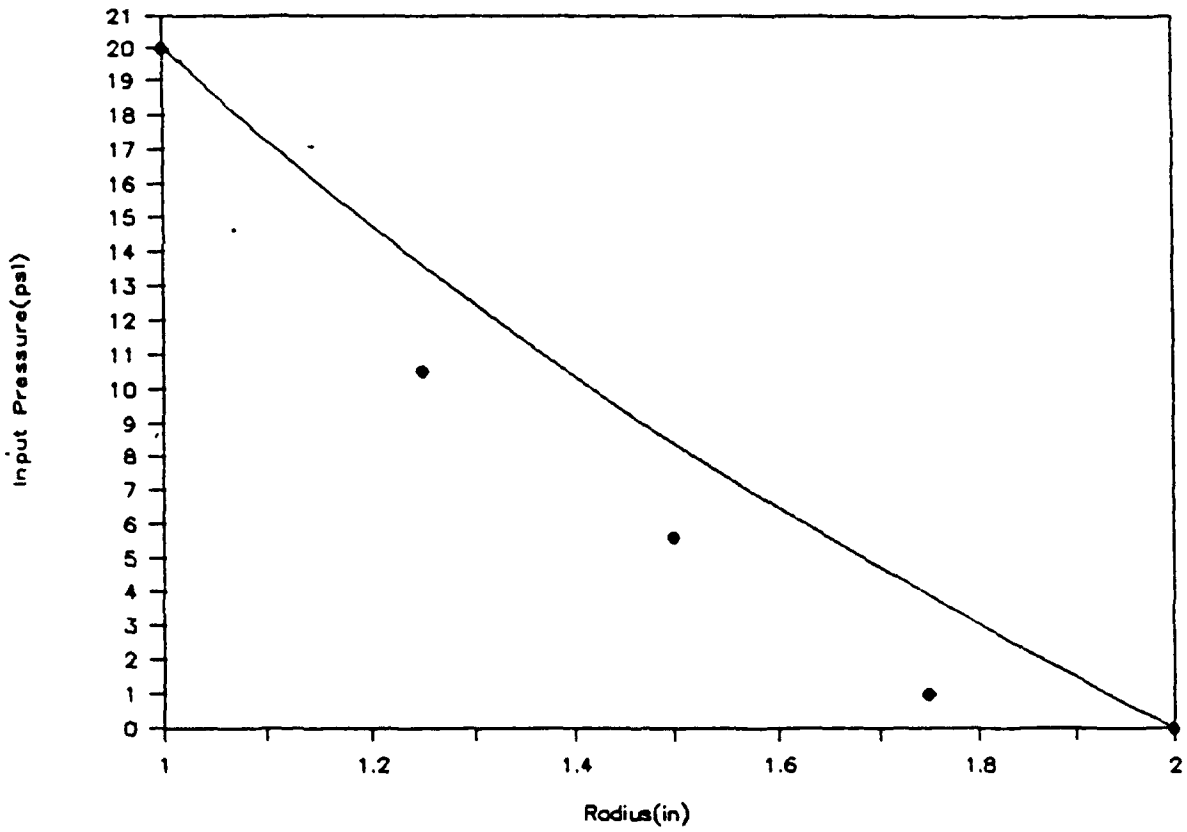


Figure 8 - Hydrostatic Pressure Profile
* 1:8 Graphite/Ethylene Glycol Mixture
* Film Thickness = 0.004 inches
* Flow Rate = 155 ml/min

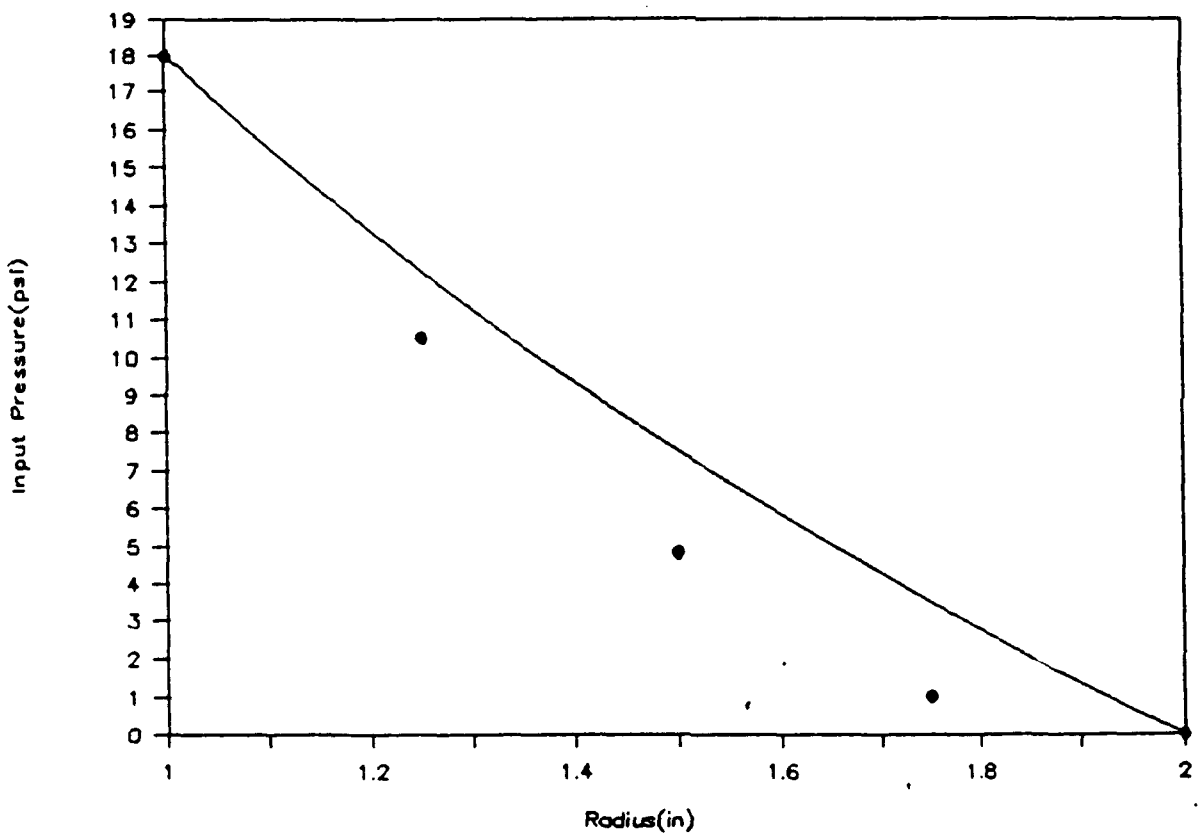


Figure 9 - Hydrostatic Pressure Distribution
- 1:8 Graphite/Ethylene Glycol Mixture
- Film Thickness = 0.005 inches
- Flow Rate = 270 ml/min

Measurements of Hydrodynamic Pressure

The experimental test rig shown in Figures A3 and A4 was used to collect the data given in this section of the report. The objective here is to collect performance data for the graphite/ethylene glycol slurry under a hydrodynamic fluid film condition. Initially, the hydrodynamic journal bearing test rig was used to collect fluid film pressure data using only pure ethylene glycol, which is the carrier fluid, as the lubricant. Figure 10 shows a schematic of a journal bearing and a typical pressure profile for a Newtonian fluid. The initial testing with strictly ethylene glycol gave baseline data from which to compare slurry pressure data.

A unique feature of the hydrodynamic test rig is the ability to control fluid film geometry. In a typical journal bearing the fluid film geometry adjusts to the externally applied load. In this case the applied load and fluid pressure profile adjusts to the film geometry. This feature allows the minimum film thickness to be varied in a controlled manner. With such an arrangement, the position the minimum clearance (or minimum film thickness) is at the bottom of the bearing and 180° from the lubricant inlet position (i.e. $\phi = 0$).

Since fluid film geometry was one of the controlled variables, it was important to assure proper alignment of the journal within the test bearing. To accomplish this, the bearing block was anchored to the test rig housing to prevent the bearing from turning or from shifting laterally due to hydrodynamic fluid film pressure. Red coloring dye was applied on the inner circumference of the bearing and the journal rotated by hand under slight bearing contact conditions to check for parallelism. Also, the journal was set concentrically and run within the bearing to confirmed proper alignment; under this condition fluid film pressure measurements were zero. One reason for slight errors in pressure measurements can be attributed to the inability to judge the preciseness of the alignment. The clearance between the journal and the bearing is adjusted by off-centering the bearing with the fine threaded mechanism shown in the photograph. A Linear Variable Differential Transformer (LVDT) is used to measure the adjustments made to minimum film thickness.

Figure 11, shows the positions of the 5 pressure taps around the test bearing. Four of the taps (taps 4,5,6, & 1) are 30° apart around the bearing circumference. The fifth pressure tap (tap 2) is located longitudinally away from the plane containing the other four pressure taps in order to determine the effect of end leakage on pressure profiles. Pressure from these five taps is transmitted to a pressure transducer by means of 1/8 inch copper tubing.

The transducer is a thin wall cylinder containing four strain gauges. The strain gauges sense the expansion of the cylinder cause by the internally applied fluid pressure. The nominal diameter of the cylinder is 2 inches and its length is 3 inches. The high sensitivity of the pressure transducer is obtained by having its wall made up of 0.05 inch brass shim stock.

For the thin walled cylinder,

$$\text{Hoop stress } \sigma = (d/2t)P$$
$$\text{Internal pressures} = 2\sigma t/d = 2E\epsilon t/d$$

where,

Modulus of Elasticity for brass, $E = 15.4 \times 10^6$ psi
Wall Thickness $t = 0.05$ in and cylinder diameter $d = 2$ in.

Strains are measured using foil type strain gages connected in a 4-arm wheatstone bridge configuration, its electrical network allowing for compensation for temperature effects as well as cancellation of signals caused by extraneous loading. A Bruel & Kjaer strain gage apparatus gives direct readout of the strains. The apparatus has a sensitivity ranging from a minimum of 100μ strain to a maximum of $30{,}000\mu$ strain. A standard compressed air supply of known pressure was used to calibrate the pressure transducer and instrumentation.

The maximum rotating speed capability of the input drive was about 6800 rpm. All testing was conducted at a speed of 3100 rpm. The journal speed was determined by using a digital tachometer. One of the aspects of the testing program was to measure amount of temperature increase in the fluid film due to journal rotational speed. A thermocouple with a resolution to read 1/10th of a degree was installed at port 3 to measure the film temperature.

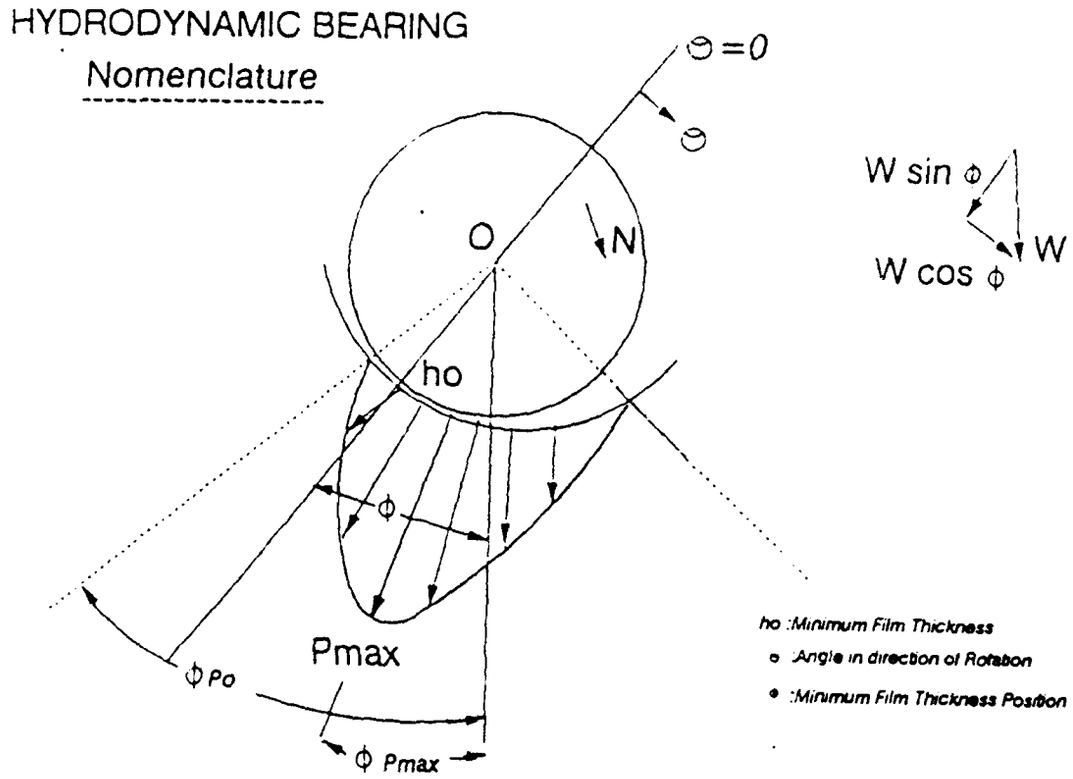


Figure 10. Schematic View of a Hydrodynamic Journal Bearing

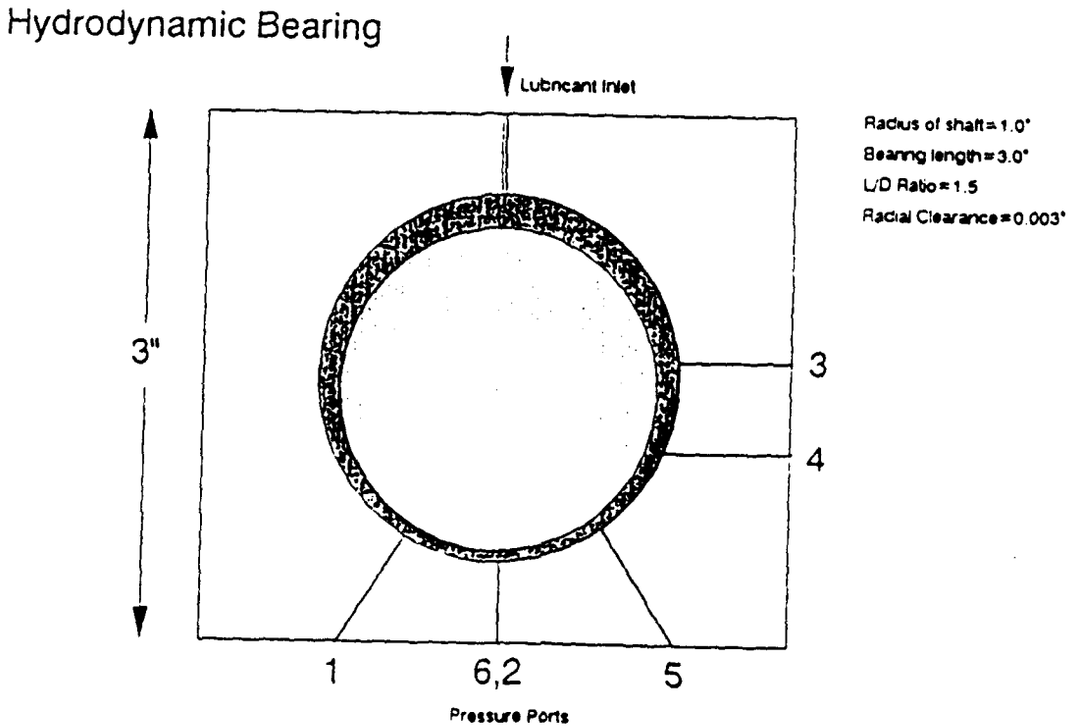


Figure 11. Bearing Showing the Position of Pressure Taps

Pressure and Temperature Measurements with Pure Ethylene Glycol

Fluid film pressures were measured at each of the 5 ports mentioned above for various film thicknesses during the course of this study. Typical pressure data is given in Table 1 for three different values of minimum thickness. A polar plot of this data follows a typical pressure profile as shown in Figure 10. Highest pressures were obtained at pressure port 5. With the pressure taps located 30 degree apart, it was difficult to estimate the precise magnitude and position of the peak pressure. These pressure readings were taken after fluid film temperature stabilized.

Table 1

Measured Pressures for Pure Ethylene Glycol
(psi)

h_{min}(in.)	Port 1	Port 6	Port 5	Port 4	Port 2
0.0015	0	5	28	7	5
0.0012	5	18	40	10	16
0.0010	6	25	49	12	24

Table 2 shows how the fluid film temperature increased as a result of fluid friction. The data indicates there was about a 15 degree Fahrenheit increase in temperature over a 25 minute period of operation. This temperature rise was large enough to cause a substantial drop in measured pressure as indicated in Table 2. This steady drop in pressure is due to a reduction in the viscosity of ethylene glycol.

Table 2

Rise in Lubricant Film Temperature During Test Run

(Pure Ethylene Glycol)

Time (min)	Pressure (Psi)	Temp (°F)
0	22	75
5	19	80
10	17.5	83
15	16	86
20	15.5	88
25	15.2	89

The next step in the study was to compare these pressure measurements with lubrication film pressures predicted by classical analytical methods.
Analytical predictions of pressure corresponding to the above test conditions are discussed below.

Theoretical Predictions of Pressure for Ethylene Glycol

D.D. Fuller (Ref. 1) summarizes the state-of-the-art for predicting lubrication film pressures. It is common practice to account for the effects of end leakage through a load reduction factor, ^{n} . Following Fuller's approach,

$$h_{min} = m\ R\ (1 - \epsilon) \tag{4}$$

For a film thickness of 0.0012 in., the corresponding eccentricity ratio of 0.6, journal speed of 3100 rpm and viscosity of ethylene glycol as 14 cP,

$$\begin{aligned} P_{avg} &= A\eta\ Z\ N\ /\ 132\ (1000\ m)^2 \\ &= 49.84\ psi \end{aligned} \tag{5}$$

Factor A = 2.12 relates to the vector representing in direction and magnitude, the resultant load carrying capacity of the film pressure. Table 3 gives the theoretically predicted pressures for other film thickness based on Fuller's equations.

Table 3

Theoretical Pressures Predicted Using Fuller's Equations

h_{min} (inches)	ϵ	P_{avg} (psi)
0.0018	0.4	21.84
0.0015	0.5	32.48
0.0012	0.6	49.84

Raimondi and Boyd (Ref. 2) have evaluated performance characteristics for full journal bearings. Their characteristic plots for

h_{min} = 0.0012 inches
ϵ = 0.6
L/D ratio = 1.5

predict

S (Sommerfeld No) = 0.0389
θ_{pmax} = 18 degrees
P_{max} = 1.397 P_{avg}
= 60 psi (at P_{avg} = 49.84 psi)

The data obtained for pure ethylene glycol although in good agreement are lower than the theoretical predictions for the same journal speed and same eccentricity ratios.
The calculations by Raimondi and Boyd are based on (a) full journal bearing having oil film extending completely around the bearing (b) no end leakage and (c) some viscosity at all point in the lubricant film. These conditions do not exist practically and pressures of a practical bearing would be less than that indicated. An attempt to make a quantitative comparison of the analysis and experiments presented is difficult for the above reasons.

Temperature has had a very vital effect on the pressures due the dependability of viscosity on temperature. The variation of viscosity with temperature for ethylene glycol is illustrated in Figure 12. It has been previously state that an approximate rise of 15° is observed during a test run.

Pressure and Temperature Measurements for the Graphite/Ethylene Glycol Slurry

The slurry mixture comprising of one part (weight) of graphite to eight parts (weight) of ethylene glycol was mixed and tested in the hydrodynamic journal bearing testing rig. Minimum film thickness was varied much the same as for the pure ethylene glycol tests. Pressure measurements for the graphite slurry are given in Table 4.

Table 4

Measured Pressures for Non-Newtonian Lubricant
(1:8 (by weight) Graphite and Ethylene Glycol Slurry)

	Pressure (psi)				
h_{min} (in.)	Port 1	Port 6	Port 5	Port 4	Port 2
0.0015	10	20	20	5	20
0.0012	25	32	32	8	30
0.0010	34	60	45	10	59

Measurements of film temperature for the slurry testing are given in Table 5.

Table 5

Rise in Lubricant Film Temperature During Test Run
(Graphite Slurry)

Pressure (Psi)	Temp (°F)
24	75
20	83
18	88
16	90
15.5	93
15	95

Variations of film temperature for the graphite slurry under the same conditions as used for pure ethylene glycol reveals that fluid friction is significantly greater when graphite is added to pure ethylene glycol. The limiting coefficient of friction for the slurry corresponds to the coefficient of graphite acting alone. Union Carbide (ref 4) gives the coefficient of friction of graphite as 0.24.

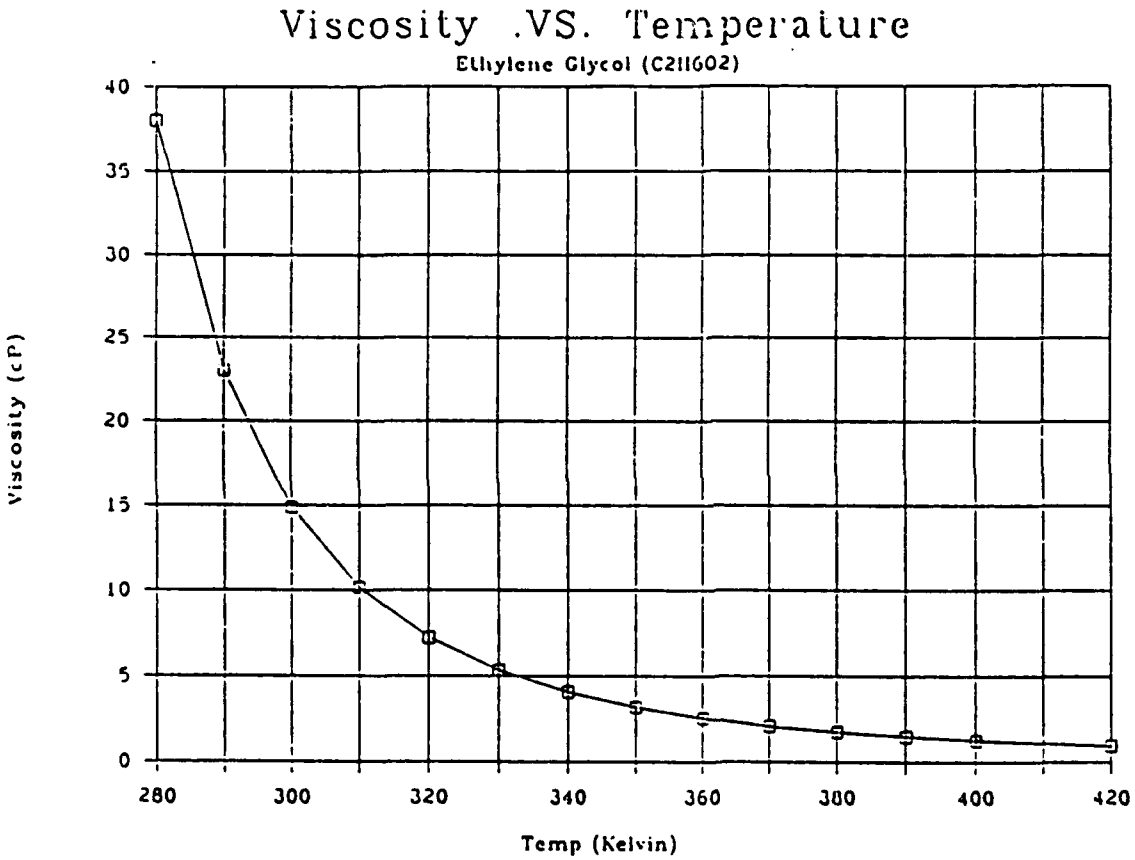


Figure 12 - Variation of Viscosity with Temperature for Pure Ethylene Glycol

CONCLUSIONS

The objective of this project was to determine the affects of adding a powder lubricant to a carrier fluid on lubrication film pressure. The effects are best evaluated in terms of the difference in fluid film pressure between the pure carrier fluid and the powder slurry. Because of the short duration of this project, only graphite was considered and the carrier or transport fluid was ethylene glycol.

Two types of lubrication films were studied. One film type was hydrostatic. The other was hydrodynamic. Two separate laboratory test rigs were designed, built and instrumented.

Test data taken with the hydrostatic test equipment showed that laboratory test data agreed with theory predictions based on the non Newtonian rheological properties of the graphite slurry. This agreement gives a high level of confidence in the use of the fluid film pressure equations derived and given in this report. As the data collected from the hydrostatic rig is based on non rotating conditions, more work is need to include the effects of friction and inertia forces in the mathematical and experimental studies.

Test data taken with the hydrodynamic test equipment was taken under dynamic or rotating conditions. The addition of graphite to the ethylene glycol produced an increase in hydrodynamic pressure above that developed by the pure ethylene glycol lubricant. The graphite increased the effective viscosity of the carrier fluid causing not only an increase in hydrodynamic pressure but also an increase in fluid temperature.

This work was meant to be a pilot study to the more direct application of powder lubricants in high temperature applications. The results, however have possible application to low temperature equipment such as squeeze film dampers and pumping losses in supplying slurries to high temperature areas.

NOMENCLATURE

h_{min}	Minimum film thickness, in.
c	Radial clearance, in.
$\epsilon = e/c$	Eccentricity ratio, dimensionless
r	Journal radius, in.
D	Journal diameter, in.
L	Axial length of bearing, in.
$m = c/r$	Clearance modulus of bearing, dimensionless
W	Load, lb.
N	Journal speed, rpm.
Z	Viscosity, centipoise
S	Sommerfeld number, dimensionless
P	Film pressure, psi.
P_{max}	Max. Pressure developed in the film, psi.
θ	Angular coordinate; measured in direction of rotation, deg.
ϕ	Position of minimum film thickness, deg.
θ_{pmax}	Position of maximum film pressure, deg.
θ_{pO}	Position at which film terminates, deg.
n	Side leakage factor

APPENDIX

Photographs of Experimental Test Rigs

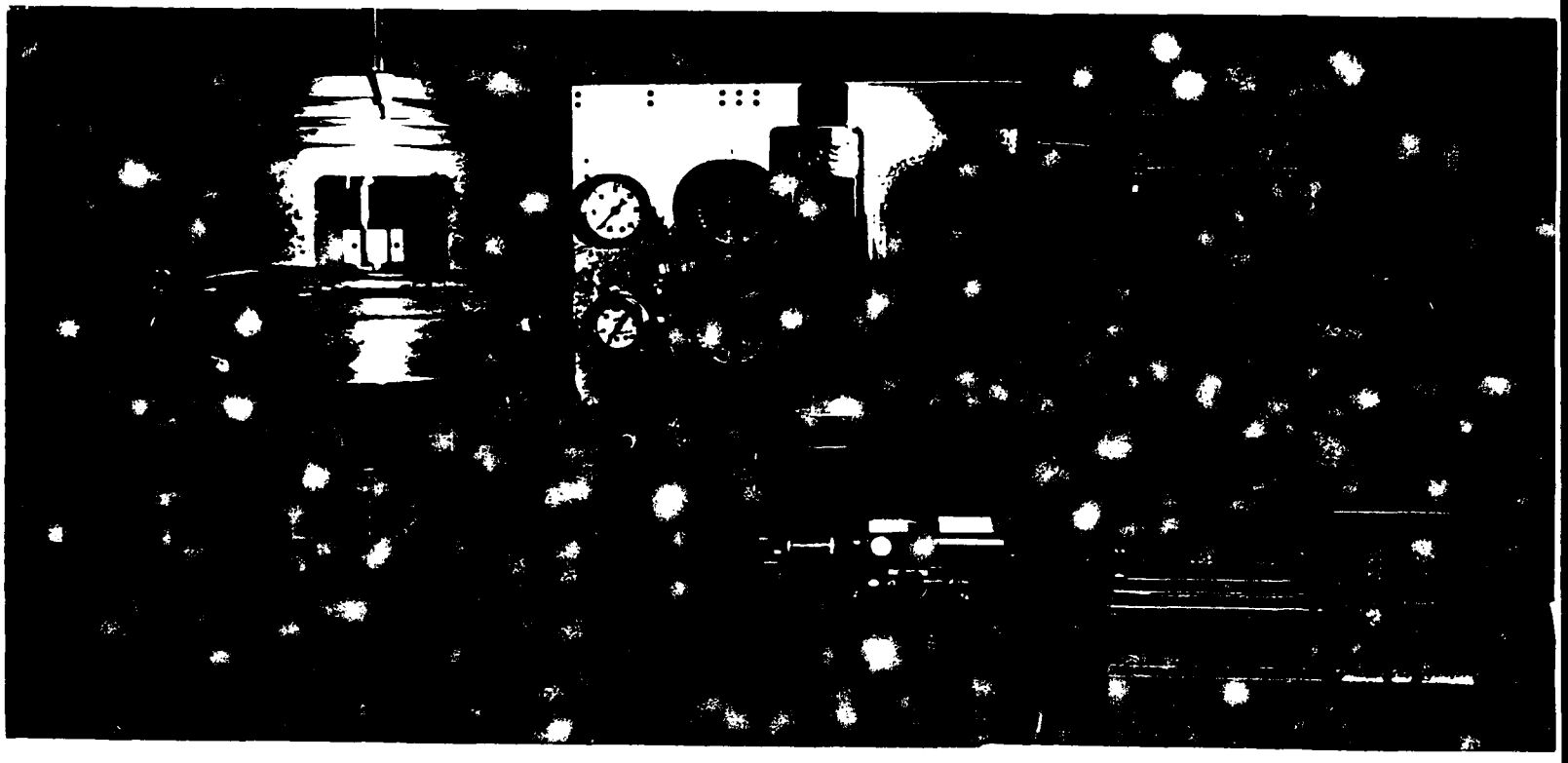


Figure A1 - Hydrostatic Experimental Test Rig

Figure A3 - Hydrodynamic Experimental Test Ri

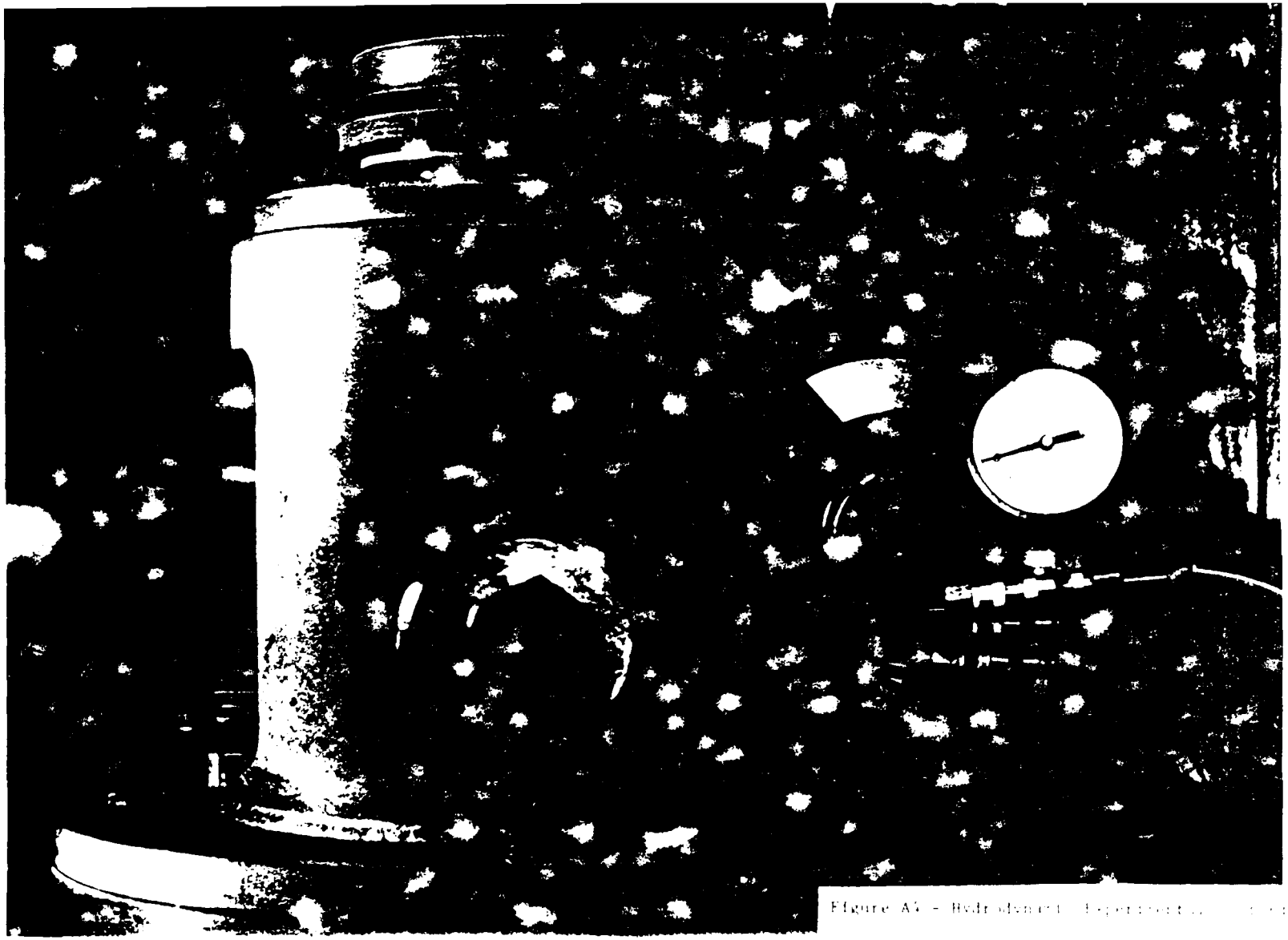


Figure A4 – Hydrodynamic [illegible]

FINAL REPORT

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

Investigation of the Combustion Characteristics of Confined Coannular Swirling Jets with a Sudden Expansion

Prepared by

David K. Pyper, B.S. Candidate

Mechanical Engineering Department

and

Paul O. Hedman, Professor

Chemical Engineering Department

Brigham Young University

Provo, Utah 84602

USAF Sponsor

W. M. Roquemore, Ph.D.

Aero Propulsion and Power Laboratory

Wright-Patterson AFB

23 December 1991

ABSTRACT

This report contains a brief summary of the work completed on a Research Initiation Grant that was sponsored by the Air Force Office of Scientific Research (AFOSR) and funded through Universal Energy Systems, Inc. (UES). The main thrust of the Research Initiation Grant was to install a burner in the Combustion Laboratory at Brigham Young University that had been designed to "specifically reproduce recirculation patterns and LBO (lean blow out) processes that occur in a real gas turbine combustor" (Sturgess, et al., 1990). Secondary objectives of the study were to conduct operational checkout tests of both burner configurations, conduct limited lean blow out experiments with the burner, and begin the design and installation of the laser based diagnostic capability that exists at BYU onto the burner.

The burner, which was provided to BYU by the Aero Propulsion and Power Laboratory at Wright Patterson Air Force Base (APPL,WPAFB), was designed by researchers at Pratt and Whitney Aircraft, Inc. (Sturgess, et al., 1990), and was fabricated in the machine shops at Wright Patterson Air Force Base. There are two configurations used in the burner. The first, which is referred to as the Pratt-Whitney Task 100 burner, uses a central fuel tube surrounded by a concentric air jet. The second configuration, which is referred to as the Pratt-Whitney Task 150 Burner, has replaced the central fuel and air tubes with an actual high swirl injector out of a commercial Pratt-Whitney jet engine.

During this study, a test facility (which includes a translating test stand, a water quenched exhaust hood, and various air and fuel feed systems for the combustor) was designed, installed, and successfully checked out. Optical systems were also designed to interface the laser based diagnostic capabilities that exist in the Chemical Engineering Department at BYU with the burner test facility. The actual interfacing of the laser based instruments will be completed under a second AFOSR Research Initation Grant that is being funded through Research and Development Laboratories, Inc. (RDL). Additional combustion experiments including CARS temperature measurements and laser sheet lighting of the combustor flames will be conducted under the new RDL Research Initiation Grant.

Limited LBO data was collected in this study with both Task 100 and Task 150 configurations. Preliminary comparisons have been made between results from the Task 100 and the Task 150 configurations operating at BYU and with data from the identical burner located at WPAFB. The values of ϕ_{LBO} for the BYU Task 100 test results were in the range of 0.72 - 0.83. Similar values for this same burner obtained at WPAFB were in the range of 0.51-0.55. Preliminary analysis of this data suggests that the reduced ambient pressure at BYU (ca 12.6 psia) compared to the ambient pressure at WPAFB (ca 14.2 psia) may be responsible for the differences in ϕ_{LBO} that have been observed. Further investigation of this effect are planned under the new RDL Research Initiation Grant.

Values of ϕ_{LBO} for the Task 150 configuration ranged from 0.45 at air flow rates of about 250 slpm to a peak of about 0.6 at air flow rates of about 350 slpm. A decrease in ϕ_{LBO} was observed as air flow was increased beyond the 350 slpm range to a value of about 0.34 at an air flow rate of about 1000 slpm. In general, the ϕ_{LBO} results for the Task 150 burner measured at BYU agreed with values of ϕ_{LBO} obtained at WPAFB during the summer of 1991 (Hedman, and Warren, 1991). However, the values of ϕ_{LBO} obtained at BYU for the high swirl fuel injector (Task 150) were found to be significantly different than the results obtained at BYU for the Task 100 burner.

The research on the Task 100 and Task 150 burners will be continued under the support of the new RDL Research Initiation Grant. This new study will continue to investigate and study the complex flow fields and flow patterns in the Task 150 with the intent of providing an in depth understanding of the complexities of the Task 150 burner and obtaining data to verify current mathematical models of the combustion process inside a turbojet engine combustor.

INTRODUCTION

BACKGROUND

This report contains a brief summary of the work completed on a Research Initiation Grant that was sponsored by the Air Force Office of Scientific Research (AFOSR) and funded through Universal Energy Systems, Inc. (UES). The main thrust of the Research Initiation Grant was to install a burner in the Combustion Laboratory at Brigham Young University that had been designed to "specifically reproduce recirculation patterns and LBO (lean blow out) processes that occur in a real gas turbine combustor" (Sturgess, et al., 1990). Secondary objectives of the study were to conduct operational checkout tests, conduct limited lean blow out experiments, and begin the design and installation of the laser based diagnostic capability that exists at BYU onto the burner. The grant from UES was received by the Department of Chemical Engineering in November 1990. Prior to receiving the grant, the principal investigator had spent the summer of 1990 participating in the Summer Faculty Research Program at the Aero Propulsion and Power Laboratory, Wright-Patterson Air Force Base (APPL, WPAFB). During this time extensive measurements were made in the combustion flame of a burner with a confined, coannular jet with a sudden expansion (Hedman, 1990). This burner, which operates on gaseous propane fuel, is known as the Pratt and Whitney Task 100 Combustor.

A schematic of the Task 100 Combustor is shown in Figure 1. The Task 100 combustor consists of coaxial jets with a 29 mm diameter central fuel jet surrounded by a 40 mm diameter annular air jet. The jets are located in the center of a 150 mm diameter duct. A sudden expansion, rearward facing bluff body, with a step height of 55 mm, is located at the exit plane of the coaxial jets.

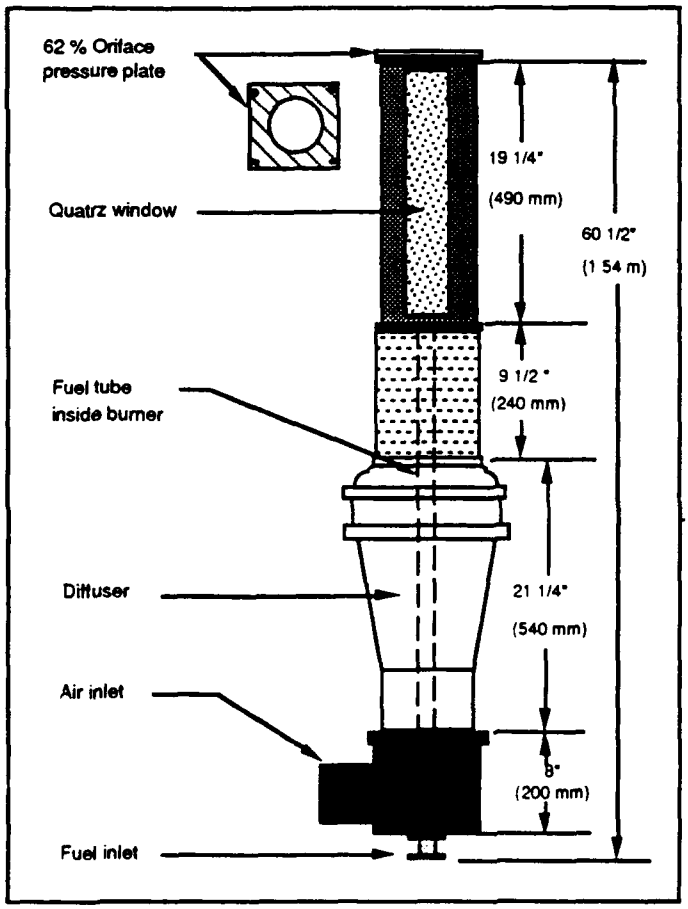


Figure 1 - P & W Task 100 Burner with 62 % Orifice Plate.

The combustor test section incorporates flat quartz windows to accommodate laser and other optical access, but uses a metal shell with metal corner fillets to reduce the vorticity and eliminate its effect on the bulk flow field in the combustor. This box-section combustor with corner fillets allows reasonable optical access, while providing a cross section that approximates a two-dimensional axisymmetric cross section. The bluff body provides a recirculation region that can stabilize the flame.

Also being studied at the Aero Propulsion and Power Laboratory, WPAFB is a rectangular combustion chamber with four swirling fuel injectors from a commercial Pratt & Whitney jet engine. This burner is known as the Task 200 combuster. It was designed to simulate a segment of a real jet engine, but with the four swirling fuel injectors very complex flow patterns and recirculation zones were produced that were extremely difficult to measure and analyze.

During the summer faculty research program of 1990 the idea of implementing a swirl fuel injector from the Task 200 burner into the Task 100 burner was conceived. This combination has become known as the Task 150 configuration The advantage of the Pratt and Whitney Task 150 Combustor was that it would allow the combustion characteristics of a jet engine fuel injector to be investigated in a simpler geometry where various diagnostic measurements (primarily laser based optical measurements) could be more easily made. A schematic drawing of the Task 150 Combustor is shown in Figure 2. A drawing that shows the installation of the high swirl fuel injector in greater detail is presented in Figure 3.

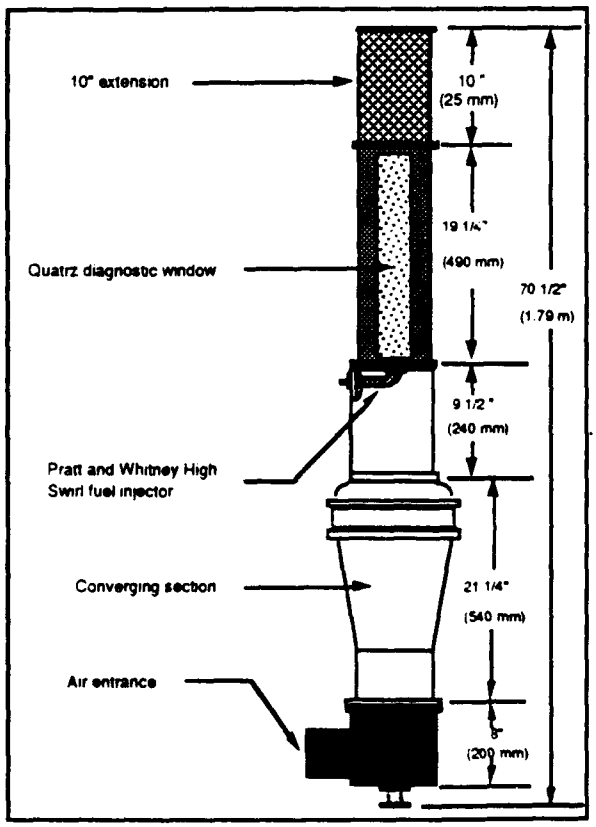


Figure 2 - P & W Task 150 Burner With High Swirl Fuel Injector.

Only a high swirl P & W swirl fuel injector has been provided for this UES (AFOSR) study. The high swirl injector is referred to by representatives of Pratt and Whitney as a "bill of materials injector" used in production engines. A low swirl injector was used at WPAFB during the 1991 summer faculty research program, but this injector was not available to this program. The Task 150 burner swirls the entering fuel and air flows as is commonly done in a real turbojet engine

combustor and could eventually allow the investigation of liquid jet engines fuels. The work in the Task 150 combustor provides a bridge between the work in the Task 100 Combustor and the Task 200 combustor.

The main thrust of this Research Initiation Grant has been to incorporate an actual swirling fuel injector like that used in the Pratt and Whitney Task 200 Combustor into the Pratt and Whitney Task 100 Combustor hardware.and examine the feasibility of making meaningful combustion diagnostic measurements in more complex swirling flows.

The burner, which was provided to BYU by the Aero Propulsion and Power Laboratory at Wright Patterson Air Force Base (APPL,WPAFB), was designed by researchers at Pratt and Whitney Aircraft, Inc. (Sturgess, et al., 1990), and was fabricated in the machine shops at Wright Patterson Air Force Base. In this study, the Aero Propulsion and Power Laboratory at Wright-

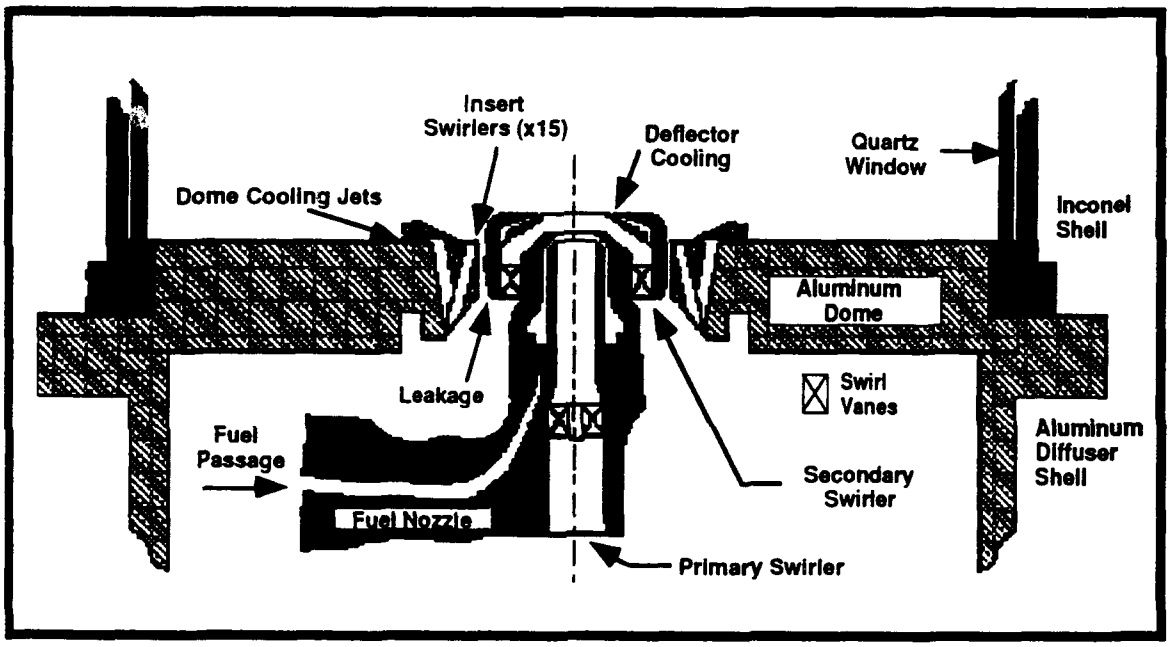


Figure 3 - Cut Away View of the Pratt and Whitney High Swirl Injector.

Patterson Air Force Base provided BYU the Pratt and Whitney Task 100 combustor hardware. Pratt and Whitney Aircraft, East Hartford, Connecticut provided a high swirl injector from an actual Pratt and Whitney turbojet engine. The Aero Propulsion and Power Laboratory began the fabrication of the Task 150 combustor the fall of 1990 and the combustor was delivered to BYU the end of March 1991. The schedule of the Research Initiation Grant was based on a delivery of hardware in November 1990. Also, the high swirl fuel injector, the aluminum dome, and the aluminum diffuser section were taken back to WPAFB during a related 1991 summer faculty research program (Hedman and Warren, 1991). The March arrival of the Task 150 hardware and

the use of some of the hardware at WPAFB during the summer of 1991 delayed the installation of the research combustor until the end of October 1991.

OBJECTIVE

The objectives of the Research Initiation Program were limited in scope and restricted to 1) the design, fabrication, and installation of a test facility for the P & W research combustor in the existing Combustion and Reactions Laboratory Building (B-41) at Brigham Young University, 2) the performance of initial checkout tests and a limited amount of lean blow out limit tests so a comparison between BYU data to previous work done at WPAFB can be performed and, 3) the initial interfacing the burner with the existing laser based diagnostics available at BYU.

APPROACH

This research program was based on the Aero Propulsion and Power Laboratory, Wright-Patterson AFB providing the burner hardware, and Pratt and Whitney Aircraft providing the Task 150 high swirl injector. The translating test stand, water quenched exhaust system, and air and fuel feed systems were to be designed, fabricated and installed by researchers at Brigham Young University. The funding of the study was rather limited. Consequently, facility components (particularly air flow system components) that were available from other facilities that have since been disassembled were used. Additionally, limited checkout and lean blow out combustion experiments were to be performed. The design of the burner facility was governed by the following constraints: 1) high temperature exhaust, 2) rigidity of a movable test stand, 3) potential for computer controlled translating test stand and data collection, and 4) limited funds.

The water quenched hood was designed to permit air and propane flow rates of up to 4000 and 104 slpm (70 F) respectively. The peak flow rate of propane gives a heat release of about 544,000 Btu/hr. The adiabatic flame temperatures near the lean flammability limit of propane and air and near stoichiometric fuel/air ratio is about is about 1508 ^{o}K (2255 ^{o}F) and 2275 ^{o}K (3400 ^{o}F) respectively (Gordon and McBride, 1976). Consequently, the water quenched exhaust hood was over designed to accommodate a heat release of one million Btu/hr and reduce the exhaust products to a temperature of about a 100 ^{o}C (212 ^{o}F) or below.

Since the combustion and flame characteristics in the burner were to be analyzed by laser techniques, the laser diagnostic volume needed to be translated in three dimensions inside the burner. The approach was to design a translating test stand that would move in two dimensions, with motion in the third dimension being provided by motion of the optical components on the laser optical tables. The translational test stand needed to be very sturdy with a minimum deflection. Also, the test stand needed to be movable. Due to physical space constraints of the building that contains the burner facility, it is necessary to move the burner out of the road so that the laser beams can be directed to other reactors located in the Combustion Laboratory. It is anticipated that the translating stand and moveable laser system components will be eventually integrated into a computer automated control system. Therefore provisions were made in the designs so as to facilitate the conversion to computer control.

The budget for the research project was reasonable yet, limited. Therefore, simplistic designs and fabrication of hardware components were needed. All of these constraints were considered during the design process of the burner facility.

TECHNICAL DISCUSSION

TEST FACILITY

The burner test facility located in building the Combustion and Reactions Laboratory Building on the Brigham Young University campus, consists of four main systems: 1) a 2 dimensional translating test stand, 2) a water cooled exhaust system, 3) the utility systems, and 4) the laser diagnostic system. The translating test stand allows the burner to be positioned vertically and horizontally such that different laser diagnostic measurements may be performed. The exhaust hood cools the exhaust from the combustor and discharges it outside of B-41. The utility system consists of three sub-systems, the water system, air system, and the fuel system. Figure 4, an over view of the burner test facility, shows a front view of the water cooled exhaust system, and a side view of the translating test stand.

2-D Translating Test Stand. The AFOSR burner is mounted on the 2-D translating test stand as shown in Figure 4. The test stand is used to interface the burner to the existing laser instruments available at BYU, and provides the translation capability to align the laser diagnostic volume inside the burner. The test stand has three main components, the translator frame, the x-axis translator, and the z-axis translator. The translator frame is constructed of

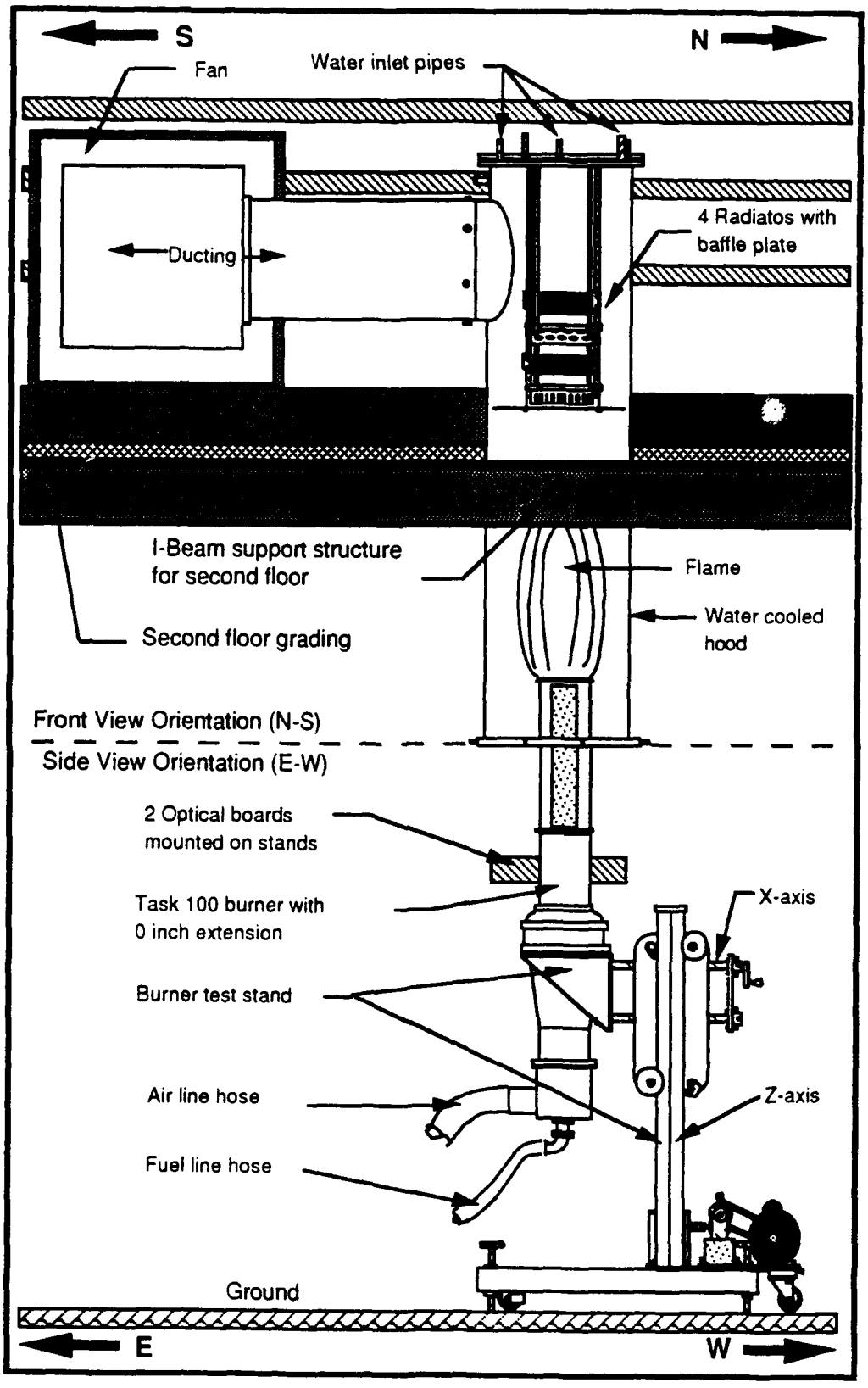


Figure 4 - Overview of the BYU (Pratt and Whitney Task 100 or Task 150) Test Facility

3 in. (76 mm) square steel tubing and a 1/2 in. (13 mm) mounting plate. The frame is mounted on four casters that allow it to be rolled about easily. The stand can be secured by positioning four leveling screws. The x-axis translator is mounted inside the z-axis carriage assembly. The x-axis translators are four hardened steel rods that slide through four linear bearings. The combustor is mounted in the burner mounting frame located at one end of the x-axis. The burner can be moved manually along the x-axis. The z-axis translator is made up of a carriage assembly that rides up or down along two square steel post. The z-axis is driven be a ball screw configuration attached to an electric motor. The translating test stand allows the diagnostic window of the burner to be positioned a maximum of 6" (150 mm) along the x-axis and a maximum of 20" (508 mm) up and down the z-axis. The third axis of movement is obtained optically through the lasers optical components and is described later in the Laser Diagnostic Design section.

Water Cooled Exhaust System. The exhaust system was designed to handle a heat release of up to one million Btu/hr. The system consists of a water cooled hood, sheet metal ducting, and an exhaust fan as depicted in Figure 4. The water cooled hood is comprised of two concentric stainless steel cylindrical shells welded at both ends in which water is allowed to enter and exit. The water enters through three inlet pipes. The pipes duct the inlet water to the bottom of the cylindrical shell. The water the flows upward through the shell where it exits out of three pipes that are manifolded together and directed to a drain pipe. The stainless steel hood is suspended from the second floor I-beams. The top of the burner can extend 8" (200 mm) up into the hood. This allows for radiation adsorption when the burner is operated in a rich high flow fuel condition.

Suspended and stacked inside the stainless steel hood is a set of four compact heat exchangers. These heat exchangers have been designed to cool the exhaust to a temperature with in the range of 180 - 200 ^{o}F (82 - 94 ^{o}C). The cooled exhaust is channeled through a set of sheet metal ducting, as shown Figure 4, and discharged outside of the building by a heat resistant fan. The heat resistant fan has a three speed variable control and pulls a maximum of 5450 CFM. Louvers were installed in the wall of the building at the fan exit to prevent backdraft into the building when the hood was not in use.

Utility Systems. The utility systems designed for the BYU burner test facility consists of three flow systems: the cooling water system, the combustion air system, and

the propane fuel system. Each of the three systems have controls and/or instrumentation devices (i.e. pressure and temperature read outs) located on a main control panel.

Water System. This system supplies cooling water to the exhaust hood and radiators. The cooling water is brought in from the building's main water supply and regulated to a water pressure of 5 psig. A thermocouple has been attached to and centered in the exhaust exit duct of the exhaust system hood to monitor the temperature of the exhaust gases. The flow rate cooling water can be regulated to insure that the temperature of the exhaust gases remains less than about 200 ^{o}F (ca 90 ^{o}C). The flow rate of the cooling water is regulated by either adjusting the on/off valve or setting the pressure regulator to a higher or lower pressure.

Air System. The air flow system was designed to control the upstream pressure to a choked flow nozzle. This provides a very simple control system since the flow through a choked flow nozzle is directly proportional to upstream pressure, In order to provide the flow ranges desired, 250 - 4000 slpm (70 ^{o}F), three separate choked flow nozzles were designed and fabricated. The throat diameters of 0.1485 in., 0.214 in., and 0.296 in., give flow rate ranges of 250-1000, 500-2000, and 1000-4000 slpm (70 F) over the operating pressure range available with the BYU air system.

The air flow control system can be operated as either a closed loop feedback control or an open loop feed control. A schematic of the closed feedback control loop is shown in Figure 5. The closed feedback control loop consists of a Val-Tek™ pneumatic control valve, a pressure gauge, a thermocouple, the set of three different choked flow control nozzles (one for each flow range), a pressure transducer, a signal amplifier, and the electronic feedback control unit.

Dry air is supplied to the burner facility from low pressure tanks located outside of the Combustion Laboratory building. The Val-Tek™ control valve regulates the pressure of the incoming air to the choked flow nozzle. The air flow rate is regulated in the closed feedback control loop by adjusting the gain control on the feedback control unit. This setting establishes the desired up stream pressure on the choked flow control nozzle. The actual upstream pressure on the control nozzle is monitored by the pressure gauge. A pressure transducer is used to provide the feedback pressure signal to the control

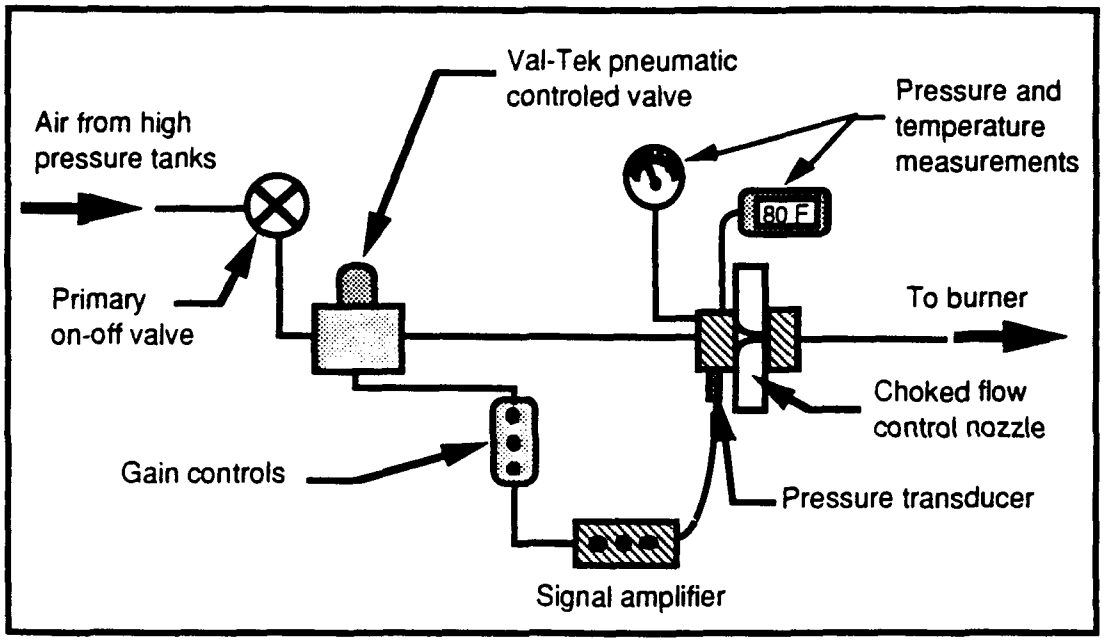


Figure 5 - Schematic of Test Facility Air Delivery System.

unit, which is amplified and compared to the gain control setting. The Val-Tek™ pneumatic control valve opens or closed accordingly. The closed loop feedback control system will hold a nearly constant pressure over an extended period of time. However, slight pressure fluctuations (ca ± 1 psig) were caused by minor oscillations in the control system. The stability of this control system is very similar to that experienced in the test facility used in Building 450 at WPAFB.

In the open loop feedback control system, the transducer, amplifier, and gain control are replaced by connecting a variable voltage supply directly to the electro-pneumatic transducer on the Val-Tek™ control valve. Since, the valve position is proportional to the voltage applied, its position can be set by adjusting the voltage. A steady voltage gives a very steady valve position, and a very steady pressure to the choked flow control nozzle. There is a slight reduction in pressure over a long time (tens of minutes) interval caused by the slight reduction in tank pressure with time as the tank pressure bleeds down. This can be easily corrected by making slight adjustments in the voltage applied to the electro-pneumatic transducer on the Val-Tek™ control valve.

The closed loop control system will be preferred when performing diagnostic measurements that require a near steady flow rate over a long time period, but where minor fluctuations in flow rate will have no appreciable effect (e.g. CARS temperature

mapping, or sheet lighting experiments). The open loop control system will be preferred when short term fluctuations cannot be tolerated (e.g. when determining lean blow out limits) and exact flow rates are critical.

In order to achieve a designed range of air flow, one of three choked flow nozzles is used. The flow rate through a given choked flow nozzle is determined from Equation 1:

$$\text{Air flow} = C_d A_t P_1 g \frac{\sqrt{[2 / (k+1)]^{(k+1)(k-1)}}}{\sqrt{gkRT_1}} \qquad (1)$$

Where C_d is the discharge coefficients, A_t is the area of throat of the nozzle, P_1 is the up stream nozzle pressure, g is the gravitational constant, T_1 is the nozzle inlet temperature, k is the specific heat ratio of air, and R is the ideal gas constant.

The initial air flow rate correlations, which are primarily a function of up stream pressure, were determined by assuming a discharge coefficients (C_d) of .98 for each of the three different diameter nozzles (eventually C_d will be determined with a flow calibration system). By using the above equation, assuming 70 ^{o}F, and using an upstream pressure range of 15-120 PSI, a flow curve has been developed for each diameter nozzle. Figure 6 presents the final correlations (Equations 2, 3, and 4) for the three air flow nozzles to be used in the burner facility.

The calibration equations obtained for each nozzle are shown below:

Air Flow Rate (slpm, 70 F)	Nozzle Diameter, in.	Nozzle Flow Rate Correlation Equation	
250 -1000	0.1485	Air Flow (slpm, 70 ^{o}F)=9.00•(P+ 12.6)	(2)
500-2000	0.214	Air Flow (slpm, 70 ^{o}F)=18.69•(P+ 12.6)	(3)
1000-4000	0.296	Air Flow (slpm, 70 ^{o}F)=35.76•(P+ 12.6)	(4)

Where P is the up stream pressure (psig) of the nozzle, and 12.6 psia is the mean ambient pressure in the BYU Combustion Laboratory.

Fuel Flow System. The fuel flow system has been designed to provide a maximum propane flow rate of 104 slpm (70 ^{o}F). Figure 7 shows a schematic of the fuel system. This system consists of a 100 lb (ca. 20 gal.) bottle of liquid propane gas with a pressure regulator, a pressure valve, a thermocouple, and two rotameters. In order to obtain high resolution measurements for both the Task 100 and the Task 150, two rotameters (same flow capacity) have been installed. One meter has a fine adjusting valve and the other rotameter has a course valve. The meters are mounted on the control

panel and are connected in parallel. The total flow rate of propane is the sum of the flow rates through both rotameters. The fine valve allows precise control of the propane flow rate through one of the rotameters, which allows precise fuel flow rates to be set over the wide range of flow rates needed for both the Task 150 and Task 100 burners.

The determination of the fuel flow rate as a function of rotameter scale reading and back pressure on the rotameter was obtained in the following manner. Air flow calibration data for both flow meters, as shown in Table 1, was provided by the manufacturer. Equation 5,

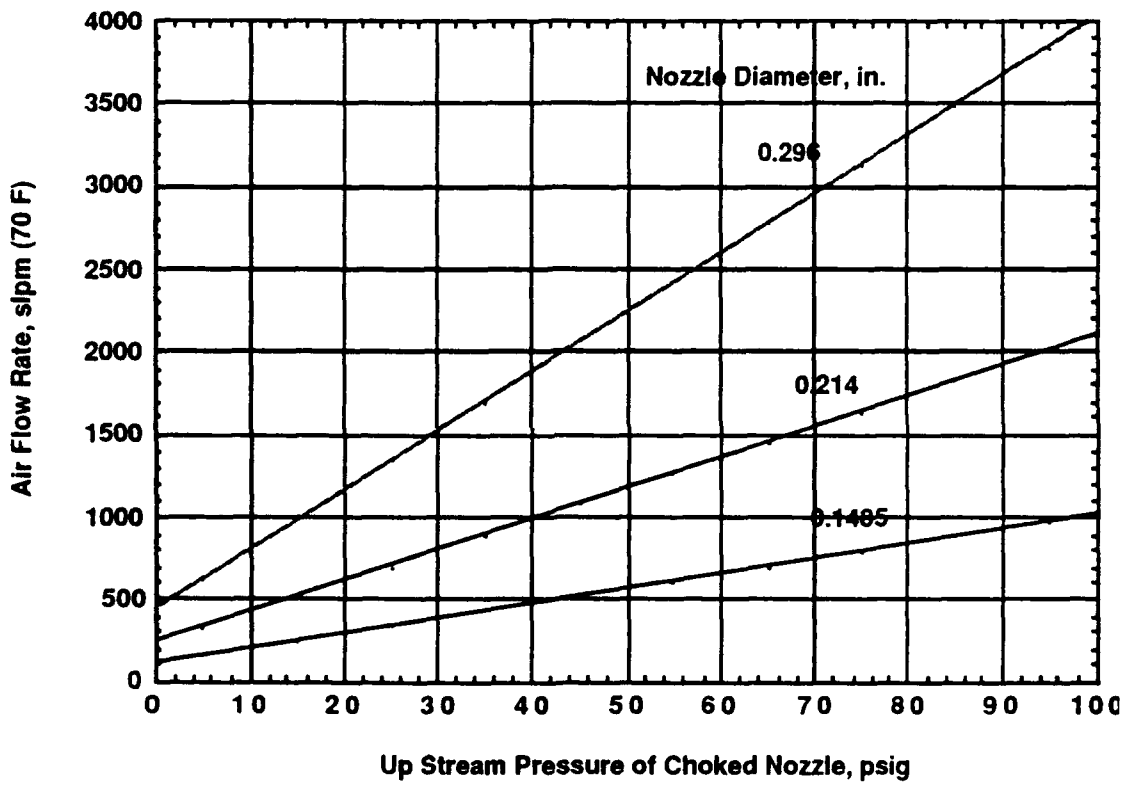


Figure 6 - BYU (P & W)Test Facility Air Control Calibration.

used for flow in a variable area flow meter (Beckwith, et al., p. 497), was used to correct the factory air flow calibration to propane:

$$Q_{C3} = Q_{air} \frac{\sqrt{(\rho_f - \rho_{C3})\rho_{air}}}{\sqrt{(\rho_f - \rho_{air})\rho_{C3}}} \qquad (5)$$

Where ρ_f is the density of the float, ρ_{C3} is the density of propane at standard conditions, and ρ_{air} is the density of air at standard conditions.

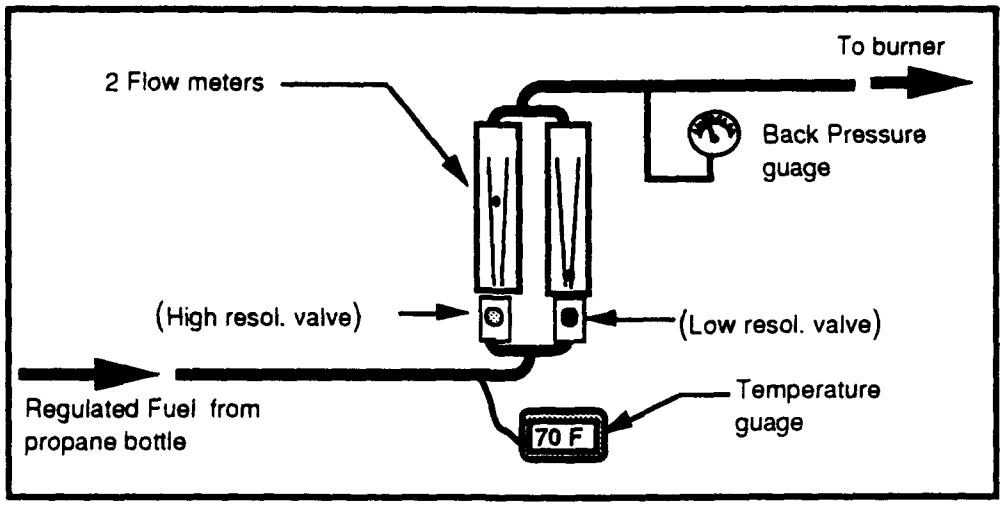


Figure 7- Schematic of Fuel Delivery System for Test Facility.

Table - 1. Manufacturer Flow meter Calibration Data for Air @ 70 F and 1 atm.

Scale Reading	Flow (slpm)
150.0	59.494
140.0	55.744
130.0	51.772
120.0	48.038
110.0	43.862
100.0	39.806
90.0	35.649
80.0	31.626
70.0	27.520
60.0	23.314
50.0	19.211
40.0	15.382
30.0	11.492
20.0	7.660
10.0	3.728

A high pressure differential is required to achieve high fuel flows through the restricted Task 150 high swirl injector. Consequently, the differential pressure (ΔP) must be considered in the flow rate correlation. The density of the propane flowing through the rotameter was modeled by the ideal gas law. A correlation of flow rate to scale reading at various upstream pressure levels was made, and the effect of upstream pressure was determined by correlating the linear curve fit coefficients (A and B in Equation 6) with pressure level. The correlations (Equations 7 and 8) for A and B were based on ambient temperature and pressure of 70 ^{o}F and 12.6 psia respectively. The correlation equations used for the propane flow rates are as follows:

$$\text{Fuel Flow (slpm 70 }^{o}\text{F)} = A + B \cdot (\text{Scale Reading}) \qquad (6)$$

Where :

$$A = -0.37013 + (1.1978\ E\text{-}2)\ \Delta P - (3.31319\ E\text{-}4)\Delta P^2 + (3.3030\ E\text{-}6)\ \Delta P^3 \qquad (7)$$

$$B = 0.34926 - (1.1303\ E\text{-}2)\ \Delta P + (2.9557\ E\text{-}4)\ \Delta P^2 - (3.1174\ E\text{-}6)\ \Delta P^3 \qquad (8)$$

ΔP is in psig

The flow rate of the propane is based on a factory air calibration table converted to propane with an assumed temperature of 70 ^{o}F. The flow rate of propane is determined from the rotameter scale reading and the down stream pressure. These correlations will eventually be checked against a calibrated flow cell.

Test Facility Operational Limits. The stoichiometric fuel to air ratio $(A/F)_{st}$ can be determined from the following stoichiometric chemical equation:

$$C_3H_8 + 5\ O_2 + 18.590\ N_2 + 0.224\ Ar + 0.007\ CO_2 \text{ ----> } 3\ CO_2 + 4\ H_2O + 18.590\ N_2 + 0.224\ Ar + 0.007\ CO_2 \quad (9)$$

The stoichiometric balance gives an $(A/F)_{st}$ equal to 23.821 (mol air / mol propane).

The value of fuel equivalence ratio (ϕ) is calculated by the following relations:

$$\phi = 23.821 \cdot \text{Fuel Flow / Air Flow} \quad (10)$$

Where:

Fuel Flow = the measured flow rate of propane, slpm (70 ^{o}F)

Air Flow = the measured flow rate of air, slpm (70 ^{o}F).

The operational envelope in terms of air flow and fuel flow for the BYU test facility are displayed in Figure 8. Lines of ϕ are shown between the rich and lean flammability limits of propane and air.

Laser Diagnostic Design. The CARS laser system is contained in a room adjacent to the burner facility. Initially, This laser system will be used to make CARS temperature measurements in both the Task 100 and Task 150 burners. The laser diagnostic system uses a Nd-Yag laser as the pump laser. This pump laser is used to provide the two pump beams needed in a folded BOXCARS phase matching configuration. The pump laser is also used to pump a dye laser that is tuned to the Stokes Raman resonance of nitrogen. The dye laser beam combined with the two pump beams in a proper phase matching configuration generates a CARS signal at the nitrogen anti-Stokes Raman frequency that contains information regarding the temperature and species concentration of the nitrogen gas. An alternate scheme that is also being considered is a USED CARS phase matching configuration. Once the CARS signal has been generated, it is returned to the spectrometer on a fiber optic cable for subsequent collection with the optical multi-channel analyzer and analysis with the Micro-VAX II computer.

The optical design for the USED CARS laser setup is presented in Figure 9. On either side of the burner are optical breadboards that are elevated to a height of 63 1/2" (1.62 m) from the ground. The third axis of movement is achieved through translation of the field lenses located near the burner windows. The diagnostic volume needs to travel only a maximum of 6" across the burner. This travel is perpendicular yet, in the same plane as the movement of the x-axis translator described above.

Other laser diagnostic techniques are being considered to investigate the combustion and flame characteristics in the burner. These techniques include LDA measurements of gas velocities, and laser sheet lighting with $TiCl_4$ seeding for flow patterns, as well as laser induced fluorescence of OH or NO using dye laser sheet lighting and film or electronic CCD cameras. The laser sheet lighting using the $TiCl_4$ seed and a film camera are readily available within the BYU Combustion Laboratory. The other laser techniques will require additional equipment that is not currently available.

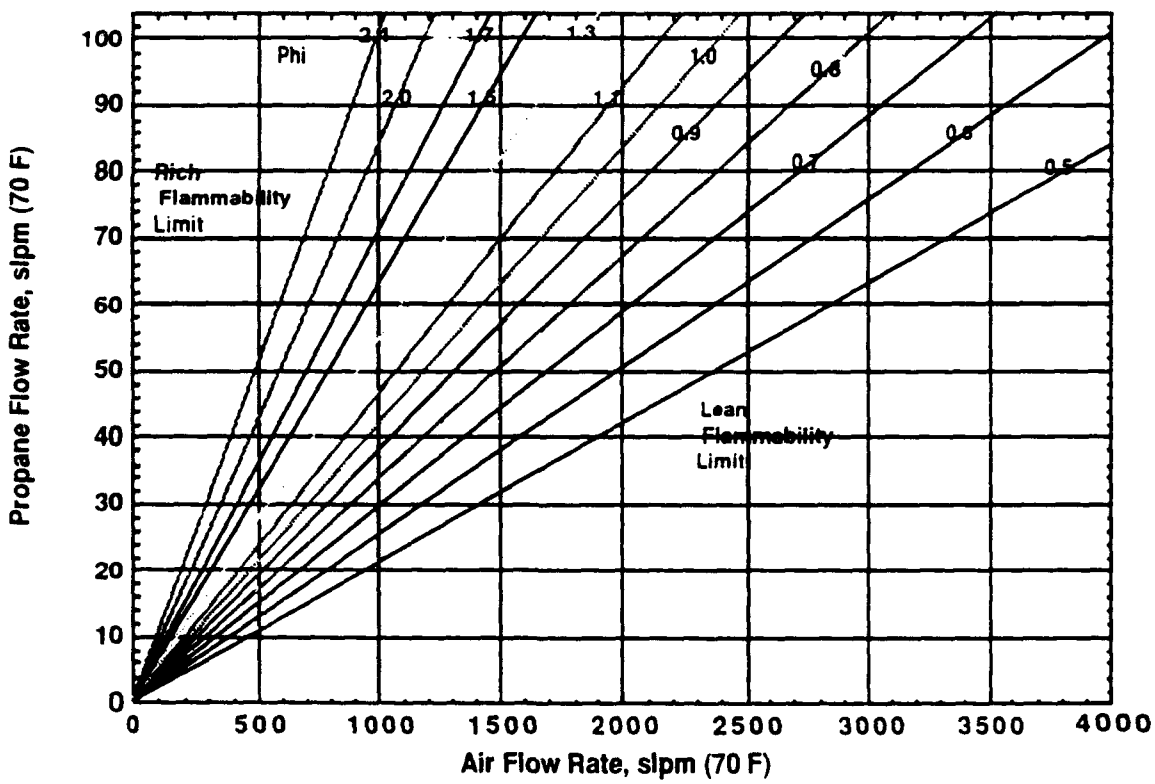


Figure 8. - Operating Limits for BYU (P & W) Test Facility.

EXPERIMENTAL RESULTS

After the all of the facility systems were assembled and installed, flow tests were performed an the air, water, and fuel systems. After correcting some minor problems, a number of

operational checkout tests were performed with the Task 150 burner configuration. The burner was found to operate well. No operational problems were found. The air flow system controlled well, and the exhaust hood cooling system proved to be more than adequate. Finally, three sets of lean blow out data were collected for with the Task 150 and the Task 100 burners. This permitted preliminary comparisons between the two burner configurations at the reduced ambient pressure in the BYU Combustion Laboratory, and to previous LBO work done at WPAFB on both the Task 100 and the Task 150 burners. A brief summary of the checkout tests and the LBO experimental results follows.

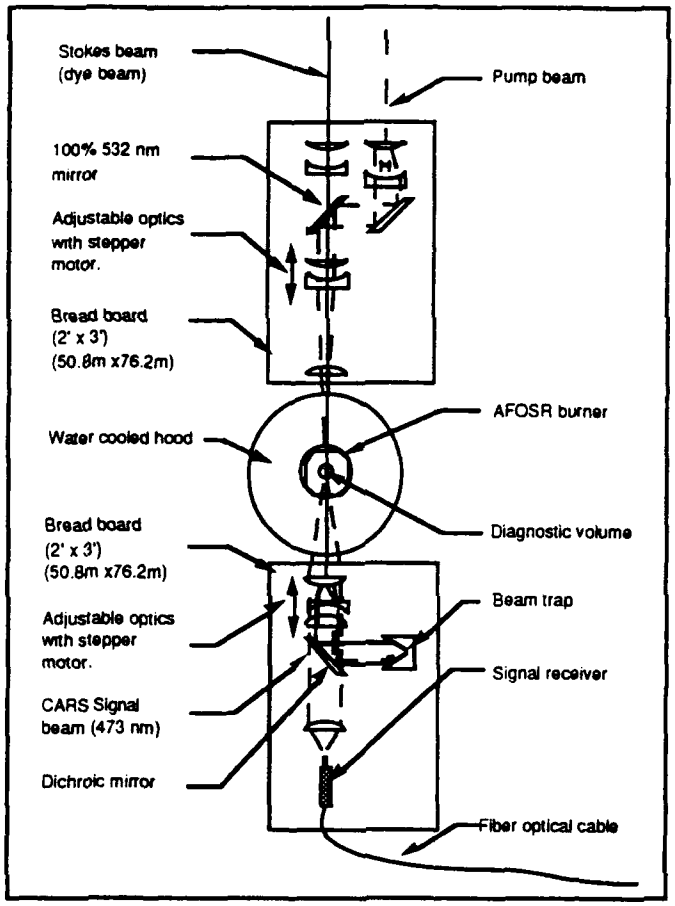


Figure 9 - Schematic of Optical Train Configuration for Used CARS

Check Out Tests. Check out tests were conducted on each of the different test facility systems to insure they operated as designed. The translational test stand positioned the burner in the x and z-axis as designed. When the water cooled hood was checked out, the inner jacket experienced some deformation due to unexpected high pressure from the main water line. The hood was repaired and a water pressure regulator set at 5 psig was installed to prevent future over pressurization. A thermocouple placed in the center of the hood exit showed that the burner exhaust was being quenched to about 200 ^{o}F (93 ^{o}C) or less. Further more, the exhaust ducting leading from the hood to the exit fan were only warm to the touch which indicated excellent cooling of the burner exhaust.

During the check out test of the air system, the closed feedback control loop could be stabilized down to a plus or minus 1 psig oscillation at 50 psig. Where as with the open feedback control loop, once a voltage was set no instability was noted in the pressure; although, the pressure did decrease slightly over long periods of time due to pressure blow down of the supply

tanks. No problems were discovered when the fuel system was checked out. Although, the needed hardware to interface the laser diagnostic techniques to the facility have been designed and purchased, the laser diagnostic system has not yet set up on the burner facility.

When Task 150 burner was checked out, it was ignited and operated through a series of air and fuel flow rates. No problems were encountered during these check out experiments.

Lean Blow Out Test Results. A series of lean blow out test were conducted with both the Task 150 and the Task 100 burners. Three sets of LBO data were obtained, one with the Task 150 burner, and two with the Task 100 burner. The Task 150 burner was configured with the 0 inch extension and 0% orifice plate for Test Set 1. For Test Set 2, the Task 100 burner was configuration the same as Test set 1. In the last test set, Test Set 3, the Task 100 burner still had no extension, but a 62 % orifice plate was was installed on the exit of the burner.

The lean blow out limit, ϕ_{LBO}, was determined by setting the air flow in the combustor, and slowly reducing the fuel flow rate (gaseous propane) until the flame extinguished. A LBO test matrix was created for each test set. The test matrix for Test Sets 1 and 2 were identical, where as Test Set 3 was similar out with fewer data points. The matrix for Test Sets 1 and 2 started at an air flow rate of 250 slpm and increased at an increment of about 50 slpm until and air flow rate of about 1000 slpm was reached. Test Set 3 began at an air flow rate of about 900 slpm and decreased in increments of about 200 slpm down to and air flow rate of about 300 slpm. Four LBO limits were collected for each air flow rate setting for all three test sets. For each test set collected, the fuel equivalence ratio at LBO was plotted against the corresponding air flow rate. Adjustments for barometric pressure differences between the several days of testing was taken into account in the air flow rate. The results for each burner are presented and compared below, and to LBO data form previous work at WPAFB (Hedman, 1990, and Hedman and Warren, 1991)

BYU Task 150 LBO Limit Results. Figure 10 presents data from the Task 150 burner. At lower air flow rates, this burner would extinguish completely from a flame that was attached to the injector. At higher air flow rates, greater than about 400 slpm, the flame would not extinguish, but would lift and stabilize on a down stream recirculation zone. Further reduction in fuel flow rate, would cause the flame to eventually extinguish completely. At air flow rates above about 740 slpm, it was never possible to get the flame to attach to the injector. Consequently, lean blow out at these higher air flow rates would always occur from a lifted flame mode. It is important to note that the attached LBO mode and the separated LBO mode overlapped in the air flow range from about 400 to about 740 slpm. Further investigation into the operational character of this burner is needed before an adequate understanding of the LBO phenomena will be obtained.

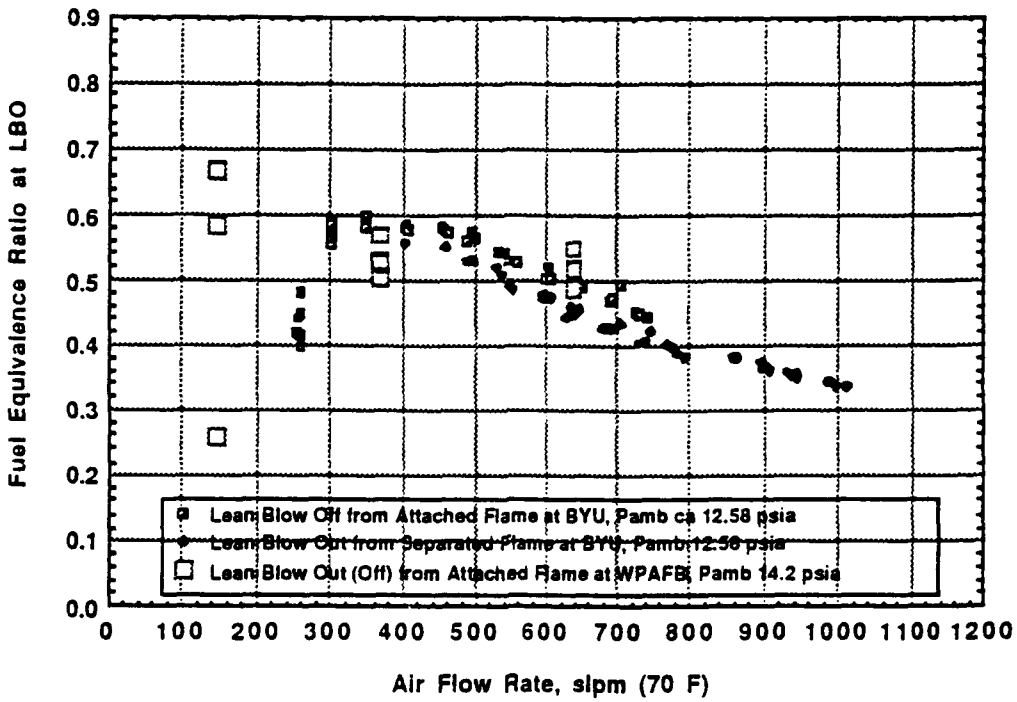


Figure 10 - Comparsion of LBO Data between WPAFB (Summer 1991) and BYU (Fall 1991) for P&W Task 150 HS Burner (0 in. Extension, 0% Orifice).

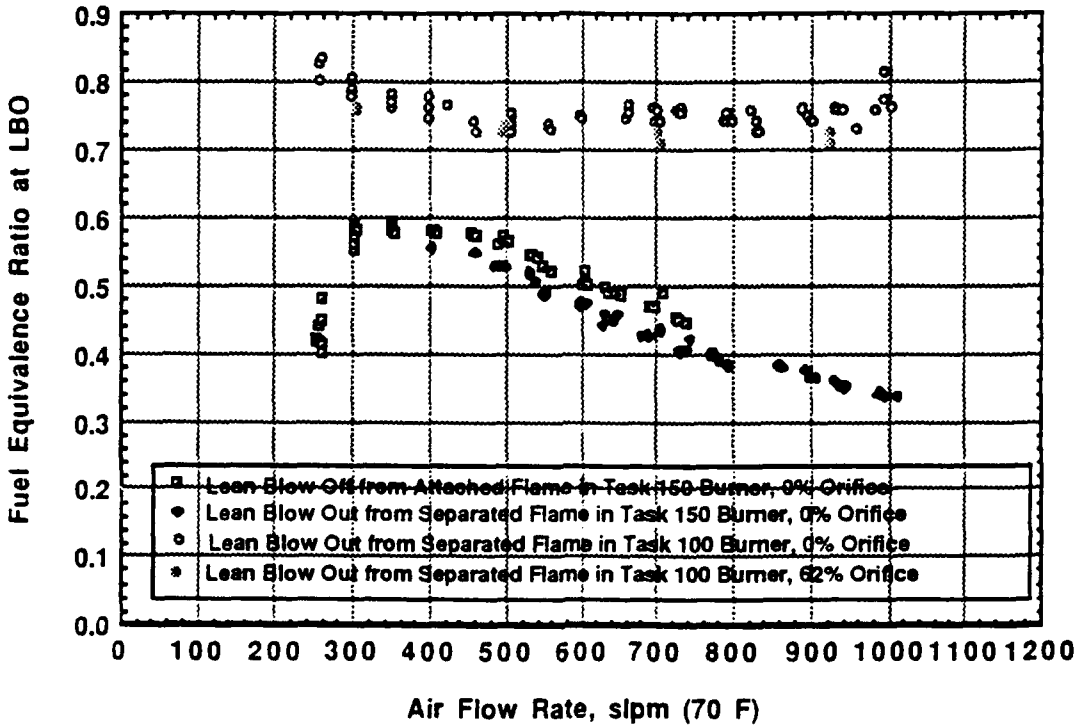


Figure 11 - Comparison of BYU LBO Data between P&W Task 150 HS Burner and P&W Task 100 Burner (0 in. Extension).

Comparison of BYU and WPAFB Task 150 LBO Data. In Figure 10, the results for the BYU Task 150 burner have been compared with results from the Task 150 burner collected at WPAFB during the summer of 1991. Both sets of data were collected under the same burnerconfiguration and operating conditions, except for a difference in barometric pressure as indicated in the figure. Since a limited amount of data was collected for this configuration and operating condition at WPAFB last summer, only a partial comparison can be made. Both sets of results are with in the same range and follow the same general trend (i.e. ϕ_{LBO} started low for both, reached a maximum and then decreased). Also, both sets of results exhibited some variation of ϕ_{LBO} during the very low air flow rates (150 -300 slpm). This variation in ϕ_{LBO} for both burners at BYU and WPAFB was thought to be caused from the very unstable flame structure that exists at those low air flow rates (i.e. low Reynolds Number). It appeared that there was reasonable agreement between the BYU Task 150 burner results and the Task 150 burner results at WPAFB. Additional data is needed in order to draw a more accurate correlation between operating the Task 150 Burner at BYU and at WPAFB.

BYU Task 100 LBO Results. No Task 100 configuration from the 1990 Summer Faculty Research Program (Hedman, 1990) exactly matched the Byu Task 100 combustor configuration or operating conditions of Test Set 2.. LBO tests at WPAFB were conducted primarily for the Task 100 burner configuration with the 10 inch extension and 0% exhaust orificeplate and with air flow rates of a 1000 slpm (70 ^{o}F) or greater. For these reasons, a direct comparison can not be made between the recent preliminary BYU data and the data taken at WPAFB. Nevertheless, a qualitative evaluation was done between the BYU Task 100 burner results and the WPAFB Task 100 burner results.

Figure 11 compares the values of fuel equivalence ratio at LBO (ϕ_{LBO}) for the Task 100 burner Test Sets 2 and 3 to the values of ϕ_{LBO} obtained for the Task 150 burner (Test Set 1). The values of ϕ_{LBO} exhibit a very different trend than do the BYU Task 150 results. The values of ϕ_{LBO} for the Task 100 burner are considerable higher than the values obtained for the BYU Task 50 burner. The Task 100 and Task 150 burners were both expected to have values of ϕ_{LBO} of about 0.50, near the lean flammability limit, similar to well-stirred reactor performance (Sturgess, et al., 1990). The Task 150 results seem to be about what would be expected, but the Task 100 results are much higher that expected, and much higher than similar results at WPAFB (Hedman, 1990). The values for ϕ_{LBO} for the WPAFB summer 1990 work were generally in the range of.50 to 0.55 and the values of ϕ_{LBO} for the BYU Task 100 results fell within a range of 0.72- 0.83. The higher values of ϕ_{LBO} for the BYU Task 100 burner compared to the from WPAFB could be attributed to the following reasons: 1) differences in the purity level of the propane fuel; 2)

incorrect calibration in the air and/or fuel flow meters; 3) the different burner configuration and air flow rates, and 4) the difference in barometric pressure at WPAFB compared to BYU.

There are several types of impurities found in propane that could increase the observed differences in ϕ_{LBO}. Inert gases like CO_2 or N_2 occupy volume and cause a lesser volume of propane to flow through the flow meters than is being metered. This could cause an apparently higher value of ϕ_{LBO}. This is thought to be an unlikely cause of the observed differences, since the values of ϕ_{LBO} for the Task 150 burners at BYU and WPAFB are in relatively good agreement.

The increased values of ϕ_{LBO} for the BYU Task 100 burner compared to that measured at WPAFB may have been caused by the air flow rate and the fuel flow rate calibrations. The coefficient of discharge (C_d) for the choked flow nozzle in the air system was assumed to be 0.98 (typical values of C_d lie with in the range of 0.97 -0.99) and the air calibration curve was determined from the formula for a choked flow nozzle. This formula gives the maximum flow rate that can possibly pass through a nozzle. Flow rates greater than predicted by this formula would have been needed in order to explain the observed results. For the fuel flow meters, the factory calibration of air flow was used and converted to a propane flow rate. This flow correlation could be in error, and could explain the observed results. However, this is also thought to unlikely since the values of ϕ_{LBO} for the Task 150 burners at BYU and WPAFB are in relatively good agreement. Since neither the air or fuel flow calibrations have been verified against a know standard (e.g. ASTM standards for choked flow nozzles) there exists some possibility of calibration errors. These flow meters will be calibrated before any final experimental results are taken or reported.

There were some differences in the two burner configurations and operating conditions. The BYU burner did not have the 10 inch extension like the WPAFB burner and the WPAFB burner was not be run at the lower air flow conditions of the BYU burner. It was not expected though that these differences would produce the dramatic differences observed for ϕ_{LBO} in the BYU Task 100 burner. Experiments will be repeated with identical geometries and air flow rates before final conclusions about the observed effects are drawn.

Although a comparison at exactly the same Task 100 geometrical configuration has not been possible (BYU has not yet received the burner extensions from WPAFB), visual observations have suggested that there are some differences in the flame structure between the flames observed in the Task 100 burners at WPAFB and BYU. The flame at BYU appeared to be somewhat longer that that observed at WPAFB. At WPAFB the ambient pressure is about 14.2 psia and the ambient pressure at BYU is about 12.6 psia, about a 1.6 psi difference.

In an attempt to understand the cause for the higher values of ϕ_{LBO} for the BYU Task 100 burner, the third test set was conducted with the 62 % orifice pressure plate installed on the burner exit. By restricting the exiting flow of the burner with the orifice plate, a slight back pressure was created in the burner. Twelve LBO limit data points were collected during this limited test set. As shown in Figure 11, the values of ϕ_{LBO} were slightly lower when compared to the values from Test Set 2. This observation suggests that the reduced back pressure in the combustion chamber when conducting experiments at BYU may have an impact on the ϕ_{LBO} results. There was no apparent effect of ambient pressure on the Task 150 experimental results. This preliminary evaluation of these results suggests that the differences in flow patterns and ϕ_{LBO} between the BYU and WPAFB experiments may be attributed to the differences in ambient pressure (12.6 psia at BYU versus 14.2 at WPAFB). Further experiments both at BYU and WPAFB with identical burner configurations are needed to fully resolve this difference. Simulation of a reduced ambient pressure with the addition of a nitrogen diluent will also be done.

ACCOMPLISHMENTS, CONCLUSION, AND RECOMMENDATIONS

ACCOMPLISHMENTS

The following have been completed during the twelve month period of the Research Initiation Grant.

1. The burner test facility was completed, operation of the Task 100 and the Task 150 has been demonstrated, and a limited series of lean blow out tests were made to check out the valid operation of the Task 100 and Task 150 combustors.

2. Lean blow-out measurements with the Task 150 configuration has been compared directly to similar measurements made at the Aero Propulsion and Power Laboratory, Wright-Patterson AFB, Dayton, Ohio.

3. A limited amount of lean blow-out measurements with the Task 100 configuration have been collected and compared to similar measurements made at the Aero Propulsion and Power Laboratory, Wright-Patterson AFB, Dayton, Ohio. Also LBO measurements with the Task 150 configuration have been collected and compared directly to the measurements from the Task 100 configuration to determine the influence of the swirling jet streams on the lean blow-out limit.

4. A design for the implementation of the existing BYU laser based combustion diagnostic instrument with the burner test facility has been completed.

CONCLUSIONS

1. The BYU LBO results for the Task 100 burner showed significant differences in ϕ_{LBO} and in the flame structures as an apparent function of ambient pressure. The values of ϕ_{LBO} obtained at an ambient pressure of about 12.6 psia were much higher than those obtained in similar work at WPAFB at an ambient pressure of about 14.2 psia. The reduced barometric pressure in the burner also caused the flame in the BYU Task 100 burner to appear longer in the combustion chamber near LBO than was observed in previous work.

2. It appeared that the high swirl injector used in the Task 150 burner prevented large changes (as was caused in the Task 100 by changes in barometric pressure) in the values of ϕ_{LBO} when compared to the results of similar work dune at WPAFB on an identical burner. Although more LBO testing with varying back pressure is needed to better substantiate this claim.

3. Two different modes of ϕ_{LBO} occur in the Task 150 in an air flow range between about 400 and 740 slpm, a ϕ_{LBO} associated with the transition of the flame from an attached mode to a lifted mode, and the ϕ_{LBO} associated with the final extinguishment of the flame from the lifted mode.

RECOMMENDATIONS

1. It is recommended that further investigations of the complex flow field and flow patterns in side the Task 150 be conducted. Flow visualization techniques could be used to examine the complex flow structures that are known to exist in the swirling and recirculating flows. Cars temperature measurements would also be important in better understanding the complex flow fields and chemical reactions.

2. It is recommended that the effects of major changes in back pressure on ϕ_{LBO} in both the Task 100 and the Task 150 be studied.

3. It is recommended that a comprehensive study be made of the transition range of the attached LBO and separated LBO limit in the Task 150. Understanding the mechanisms that governed the transition between these two operating modes is important to the design of the swirl fuel injector and combustion chamber. A better understanding of this phenomena may lead to improved combustion stability margins for jet engines.

4. It is recommended that the air and the fuel flow meters be calibrated against a known calibration standard and that the actual composition of the propane being used at both BYU and WPAFB be determined.

REFERENCES

Beckwith, Thomas G., N. Lewis Buck, Roy D. Marangoni. Mechanical Measurements, 3rd ed., Massachusetts: Addison-Wesley Publishing Company, 1982.

Gordon, S., and McBride, B. J., "Computer Program for Calculation of Complex Chemical Equilibrium Compositions, Rocket Performance, Incident and Reflected Shocks, and Chapman-Jouguet Detonations," NASA SP-273, Interim Revision (March 1976)

Hedman, P.O., "Investigation of the Combustion Characteristics of Swirled Injectors in a Confined Coannular System with a Sudden Expansion," Final Report, UES Summer Faculty Research Program (1990)

Hedman, P.O., and Warren, D.L., "Investigation of the Combustion Characteristics of Swirled Injectors in a Confined Coannular System with a Sudden Expansion," Final Report, RDL Summer Faculty Research Program (1991)

Sturgess,, G.J., D.G. Sloan, A.L. Lesmerises, S.P. Henneghan and D.R. Ballal. "Design and Development of a Research Combustor for Lean blow out Studies", 35th International Gas Turbine and Aeroengine Congress and Exposition. Brussels, Belgium, (June 1990).

Roquemore, W. M., V. K. Reddy, P. O. Hedman, M. E. Post, T. H. Chen, L. P. Gross, D. Trump, V. Vilimpoc, G. J. Sturgess. "Experimental and Theoretical Studies in a Gas-Fueled Research Combustor", AIAA 29th Aerospace Science Meeting, (7-10 January 1991).

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the Universal Energy Systems, Inc.

FINAL REPORT

AIRCRAFT HVDC POWER SYSTEM - STABILITY ANALYSIS

Prepared by:	K. Sankara Rao, Ph.D.
Academic Rank:	Professor
Department and	Electrical and Electronics Engineering Department
University:	North Dakota State University, Fargo, North Dakota
Date:	January 9, 1992

AIRCRAFT POWER SYSTEMS - STABILITY

by

K. Sankara Rao

ABSTRACT

Analysis and modeling of aircraft 270V dc electrical power systems are the main topic of the research project. HVDC at 270 volts has many advantages over the currently used three phase electrical power systems in an aircraft. There are some problems, particularly instability in the presence of a constant power load, which are addressed in this research. Computer models have been developed for the various components of the HVDC system and constant power load. The analysis using EMTP is included in this report.

Acknowledgment

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. Universal Energy Systems must be acknowledged for their help and concern in all administrative aspects of this program. I wish to thank the staff at the Electrical and Electronics Engineering Department of North Dakota State University for their help in preparing this report.

1 INTRODUCTION:

Aircraft Electrical Power Systems, at present use three phase AC systems and research and development into the possible use of 270V dc as an alternative is underway. Electrical Laboratory of Aero Power Propulsion Laboratory at Wright Patterson Air Force Base is very much concerned with computing modeling of HVDC systems. One of the main problems foreseen is instability when constant power loads are applied to the HVDC distribution system.

2 OBJECTIVES:

Electrical load on an aircraft power system usually consists of a combination of lighting and motor loads. In addition there is a considerable amount of dc load in an aircraft for flight critical computers etc. The main power supply in an aircraft at present is a three phase supply at 400 Hz. This frequency must remain constant in the presence of variable speed turbines which supply mechanical power to the electrical generators. There are two methods generally used for obtaining constant frequency.

1. Constant Speed Constant Frequency Systems (CSCF): A Constant Speed Drive (CSD) is used to keep the generator running at a constant speed even while the input to the constant speed drive varies considerably. The advantage is that the electrical output from the alternator is at a constant frequency with very little harmonic content. The main disadvantage of this system is the very high maintenance required of the CSD.

2. Variable Speed Constant Frequency Systems (VSCF): In this system a variable speed alternator supplies a variable frequency three phase power. This output is rectified and inverted to produce constant frequency three phase power. The main advantage of this system is low maintenance cost. The major disadvantage is the presence of harmonics in the output due to the inverter operation.

As can be seen above, the three phase ac distribution system has some drawbacks and research and development is under way for using a 270V dc supply as the main electrical power supply. In this system, the inverter portion of the VSCF system can be eliminated and all of the dc loads can be directly applied to it. The low voltage dc loads at 28V can be supplied by using dc-dc converters. As far as motor loads are concerned, inverter fed brushless dc motors, inverter fed induction motors or inverter fed switch reluctance motors can be used.

The Electromagnetic Transients Program (EMTP), which was developed in the early seventies for Bonneville Power Administration (BPA), is an excellent tool for analyzing transient behavior of a power system. This package, used in the present research, is constantly being revised and modified. It has been found that constant power loads cause instability in HVDC systems and use of a large capacitor and harmonic filters eliminate the instability and reduce the ripple current drawn from the power system.

A more detailed analysis of the HVDC system when motor loads are present should be undertaken. Since EMTP can be used to model various types of motors, it is possible to analyze the system under motor loads using EMTP.

3 EMTP

The Electromagnetic Transients Program was developed in the early seventies by Dr. W. Scott Meyer of Bonneville Power Administration. This program is being constantly updated and has numerous users all over the world. The program is very extensive and has all the features necessary for the analysis of Aircraft Power Systems. Some of the highlights of the program are as follows:

- The various kinds of elements that can be represented are:

1. RLC branches
2. Distributed lines
3. Switches including diodes, thyristors, power transistors and time controlled switches.
4. Sources
   - (a) Voltage sources, both ac and dc
   - (b) Current sources both ac and dc
   - (c) Voltage and Current sources controlled by any other variable
   - (d) Alternators with their full representation
   - (e) DC generators and motors with their full representation

(f) Induction motors with their full representation

(g) Analytical sources

5. Control System Blocks. These blocks can be linear or nonlinear. They can be represented by transfer function blocks for linear components. The inputs and outputs of the control system blocks can be interfaced with the voltages and currents of the electrical network.

- The output results are conveniently obtained in graphical form on a CRT or on any type of plotter or printer. In addition, there is a tabular output which provides highlights of all the necessary results.

4 CONSTANT POWER LOAD ON AN IDEAL HVDC SYSTEM

As a starting point for this research, an ideal 270V dc source is chosen as the power system. The series impedance with the ideal source is chosen as a small inductance in series with a parallel combination of an inductance and a resistance. The series inductance represents the subtransient reactance of the alternator in the actual HVDC generating system. The parallel combination of resistance and inductance represent the time constant and the open loop error. The rectifier filter is represented by a series RC circuit across the load. The constant power load is represented by a voltage controlled nonlinear current source such that the product of the current and the voltage equals the power. The constant power load is a cyclic load occurring 100 times a second with a duty cycle of 0.5. The

circuit diagram is shown in Fig. 1. The simulation results are shown in Fig. 2. Some of the main points of the results are the following.

1. There is considerable ringing in the voltage across the load.
2. The current drawn from the source has a very high ripple content with a fundamental frequency of 100 Hz.

An analysis of the circuit shows that the system is unstable for constant power loads. The system can be stabilized by including a large capacitor across the constant power load. The size of the capacitor depends on the magnitude of the load, the value of the inductance in series with the source and the resistance in the circuit. The simulation results, with a 3000 microfarad capacitor in parallel with the load, are given in Fig. 3. The ringing in the voltage is now absent but the ripple content of the source current is still high.

The ripple current can be reduced by inserting series LC circuits with resonant frequencies equal to 100 Hz and its odd harmonics, in parallel with the load. Fig. 4 shows the effect of inserting these harmonic filters.

5 SIMULATION OF A RECTIFIER WITH A CONSTANT POWER LOAD

The second phase of the research consisted of simulating an ideal three phase source rectified and feeding a constant power load cycling 100 times per second. The circuit diagram for this system is shown in Fig. 5 and the results are shown in Fig. 6. The following conclusions can be drawn from the results.

1. The constant power load causes instability and this instability can be eliminated by inserting a large capacitor in parallel with the load. The ripple content in the load voltage is still high. The current drawn has a high ripple content of 100 Hz and its harmonics.

2. Introduction of harmonic filters as suggested in the earlier section reduces the ripple content of the voltage and the current drawn from the supply.

In the next phase of the simulation, the three phase voltage source magnitude is controlled so as to have the voltage across the load to be 270 volts. The overall control system is of type one.

Results show that the instability can be eliminated by inserting a large capacitor.

6 BRUSHLESS DC GENERATOR WITH A CONSTANT POWER LOAD

In the next phase of the research, the stability of a brushless DC generator in the presence of constant power load is studied. As in the earlier cases, the overall system is unstable under a constant power load. The instability is due to the fact that the constant power load appears as a negative resistance and the overall impedance as seen by the generator has a negative real part. Introduction of a large capacitor in parallel with the load makes the overall impedance to be stable and the overall system is stable. The value of the capacitor needed to make the

system stable depends on the level of constant power that is applied as the load. The circuit diagram of the system is shown in Fig. 7 and the results are shown in Fig. 8 and 9.

7 CONSTANT POWER LOAD

In all phases of the earlier research it was assumed that the load has a characteristic of demanding constant power irrespective of the voltage across it. This load is simulated as a current source whose magnitude is such that the product of the source current and the voltage across it is the negative of the power demand.

In actual practice, the constant load, as it is assumed is a power conditioner whose output is maintained constant irrespective of the input voltage. When this power conditioner is connected to a resistive load, which is constant the output power is constant. In this phase of the research a buck switching type of regulator is simulated and used to replace the current sources. The overall system stability is studied. The circuit diagram of the regulator is shown in Fig. 10 and the results are shown in Fig. 11.

An examination of the results shows that any constant power load, which is derived from a switching regulator does not pose any problems at all when the input voltage is derived from ideal sources. The output voltage of the regulator remains constant while the input voltage is varied within wide limits. The settling time of the transient response, when the load resistance is changed, is of the order of a tenth of a millisecond. This

makes the circuit fully capable of operating a pulse load at 100 cycles per second.

8 MOTOR LOADS

The preliminary study shows that a constant power load, when viewed as a resistor with a hyperbolic v-i characteristic, poses a stability problem for dc systems. However, when the constant power load is considered as a constant resistive load on a power conditioner, the dc system seems to behave normally.

The power conditioner circuit mentioned in the earlier section is simulated with constant power load and with additional motor loads. Two types of motor load have been simulated: Brushless dc motors and inverter fed induction motors. The results are satisfactory and the dc voltage at the terminals is maintained fairly constant.

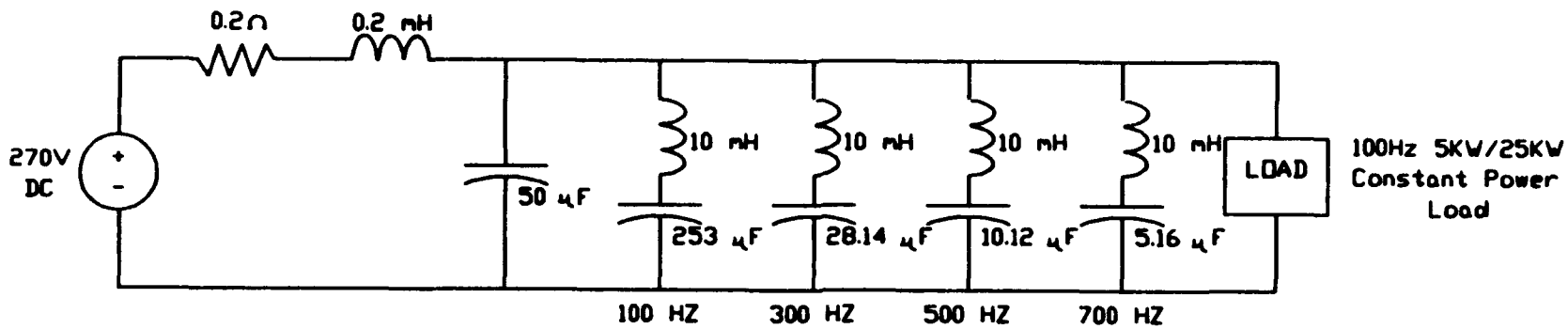
The simulations using both brushless dc motors and inverter fed induction motors show that as the motors' loads are increased, the effect of constant power load is minimized.

This research shows that a 270V dc power supply is a feasible alternative for Aircraft power supply. The istability that is asoociated with constant power loads in preliminary analysis and simulation is in fact not present when the constant power load is derived through the use of a power conditioner. In addition, the loads on the dc power supply excluding this constant power load will ensure that there is no stability problem present.

There are some additional problems associated with dc power supply. One of the main problem areas is corona. Since the voltage never crosses zero, any corona discharge once started cannot be estinguished.

9. RECOMMENDATIONS

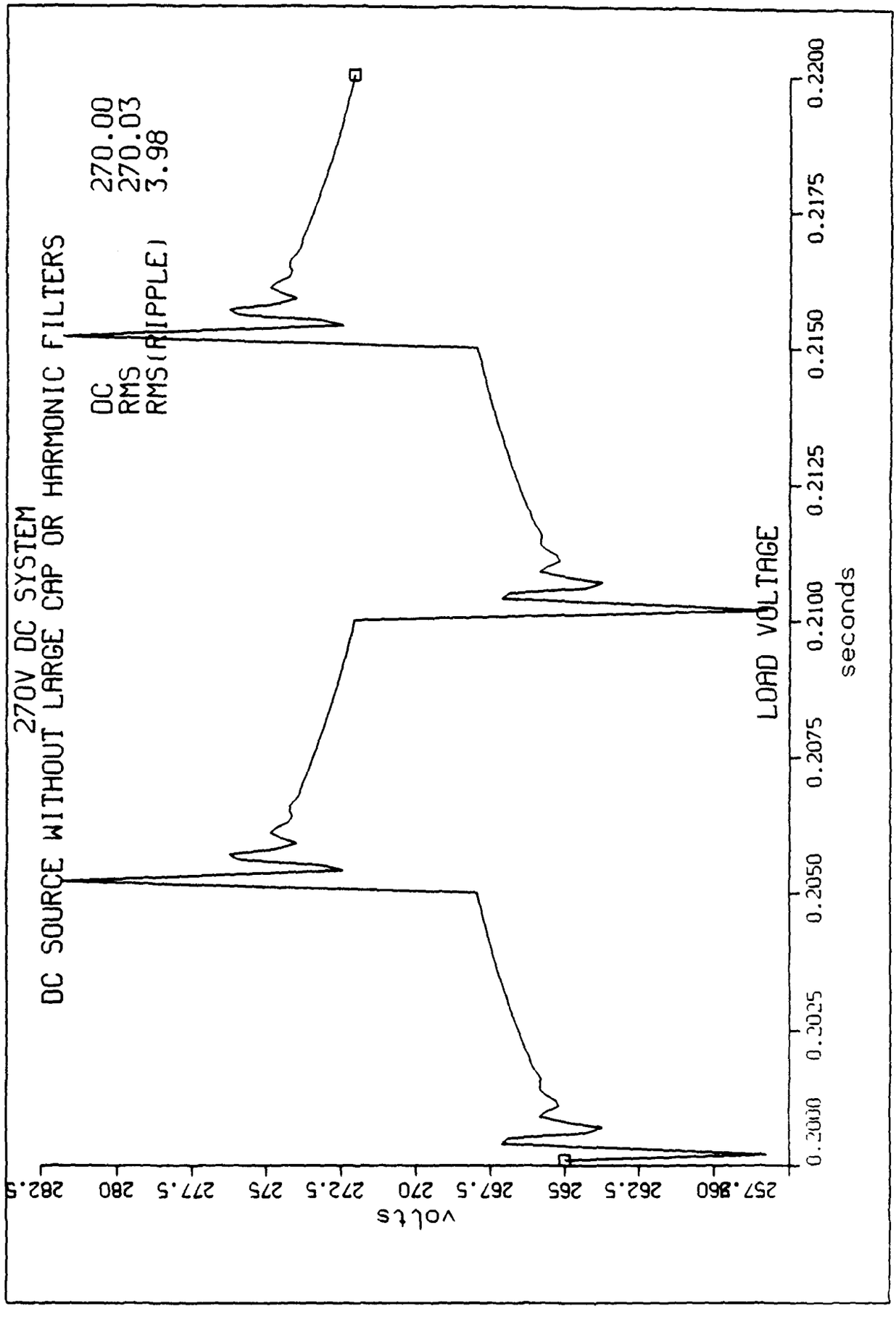
The Electromagnetic Transients Program (EMTP) a very powerful simulation tool for power system engineers has been used for the simulation and analysis of 270 V dc system when subjected to constant power loads. It is recommended that as a followup these circuits can be built and tested and the deviayin of laboratory results from the simulation should be examined. It is recommended that the circuits which have been simulated are actually built using MCT's and tested in the laboratory. This principal investigator will be submitting a proposal for doing the experimental analysis at North Dakota State University.



270V DC SOURCE

CONSTANT POWER LOAD AND HARMONIC FILTERS

Fig. 1


270V DC SYSTEM
DC SOURCE WITHOUT LARGE CAP OR HARMONIC FILTERS
DC 270.00
RMS 270.03
RMS(RIPPLE) 3.98
LOAD VOLTAGE
seconds
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200
volts
257.5
260
262.5
265
267.5
270
272.5
275
277.5
280
282.5

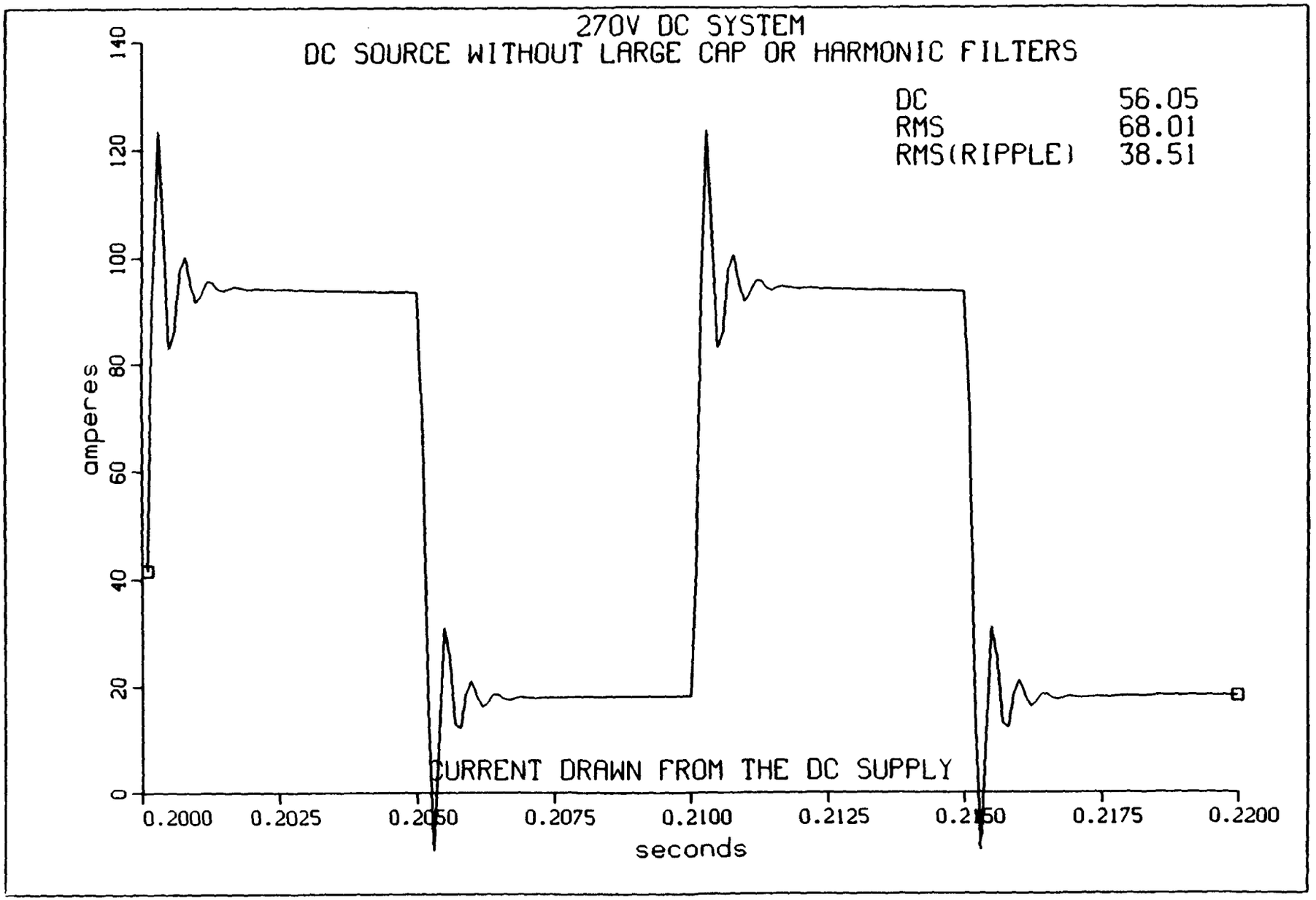
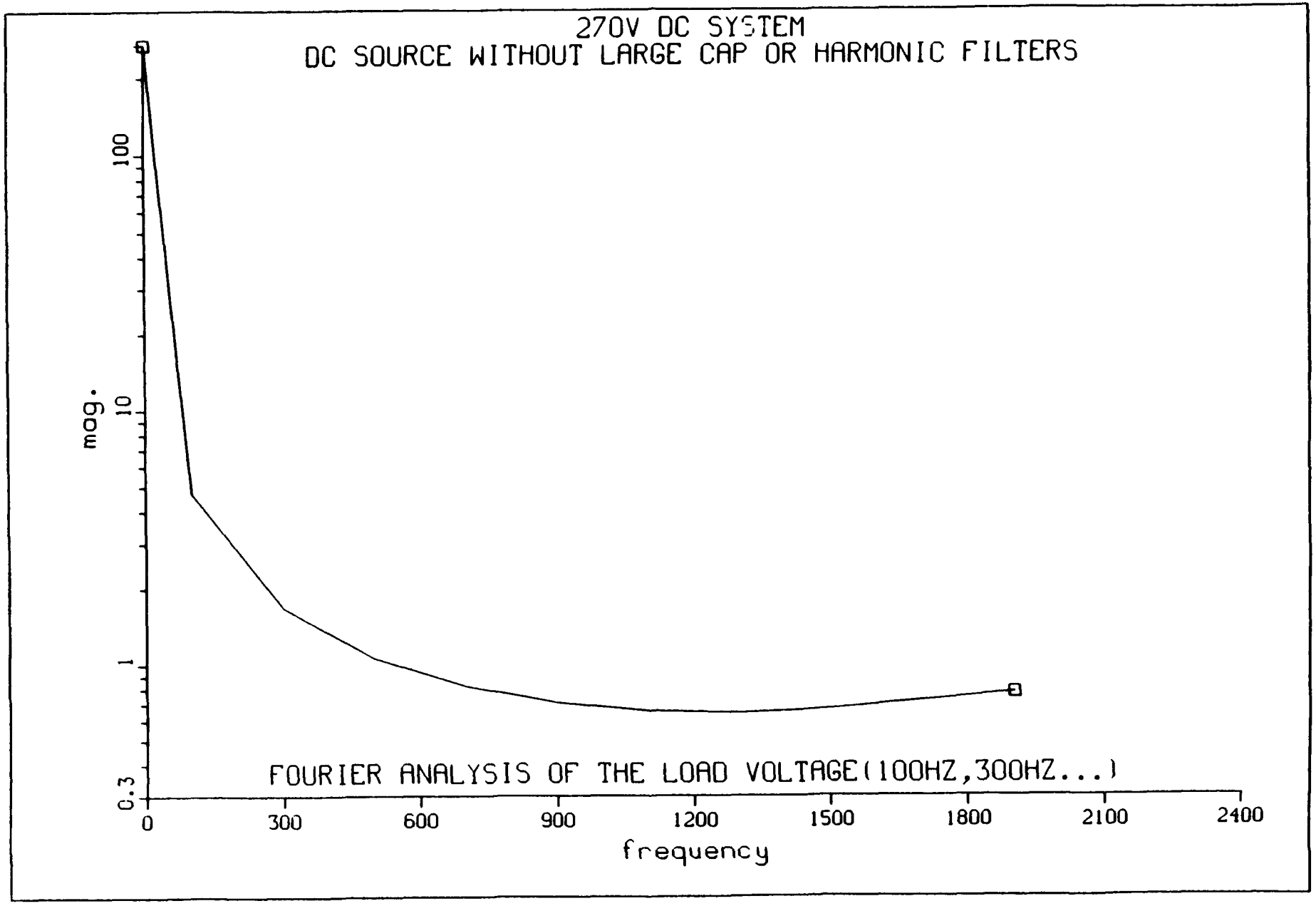
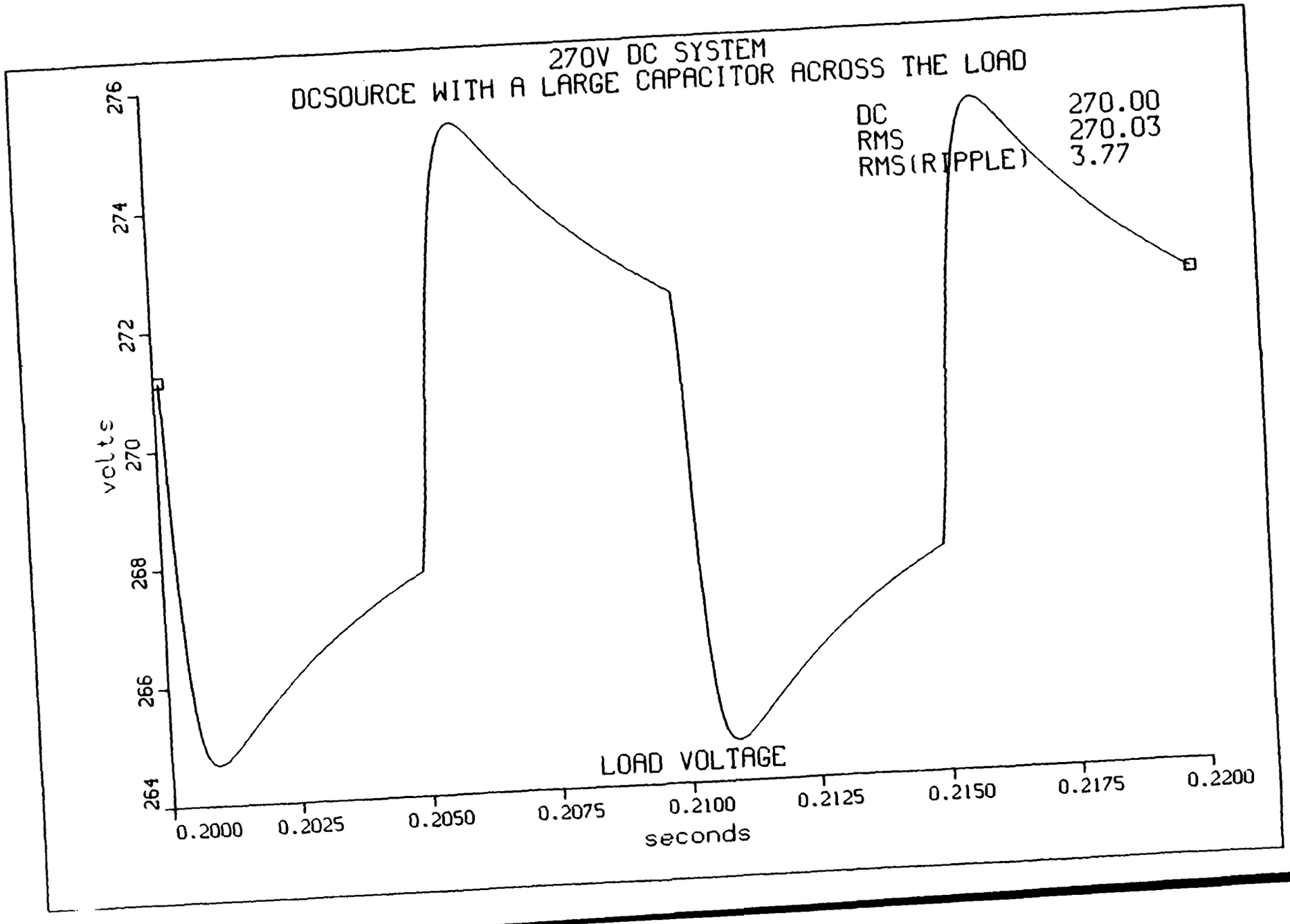
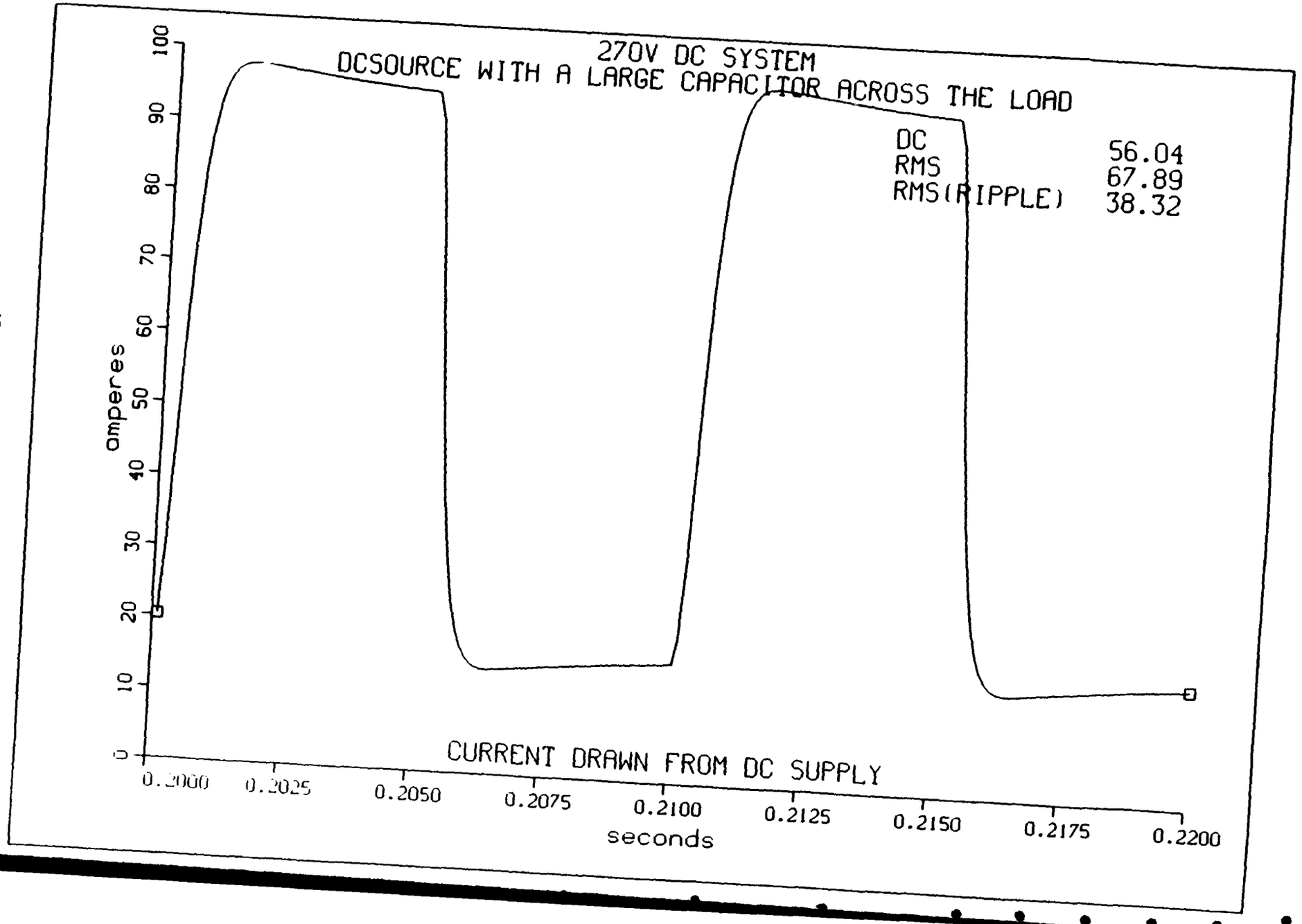
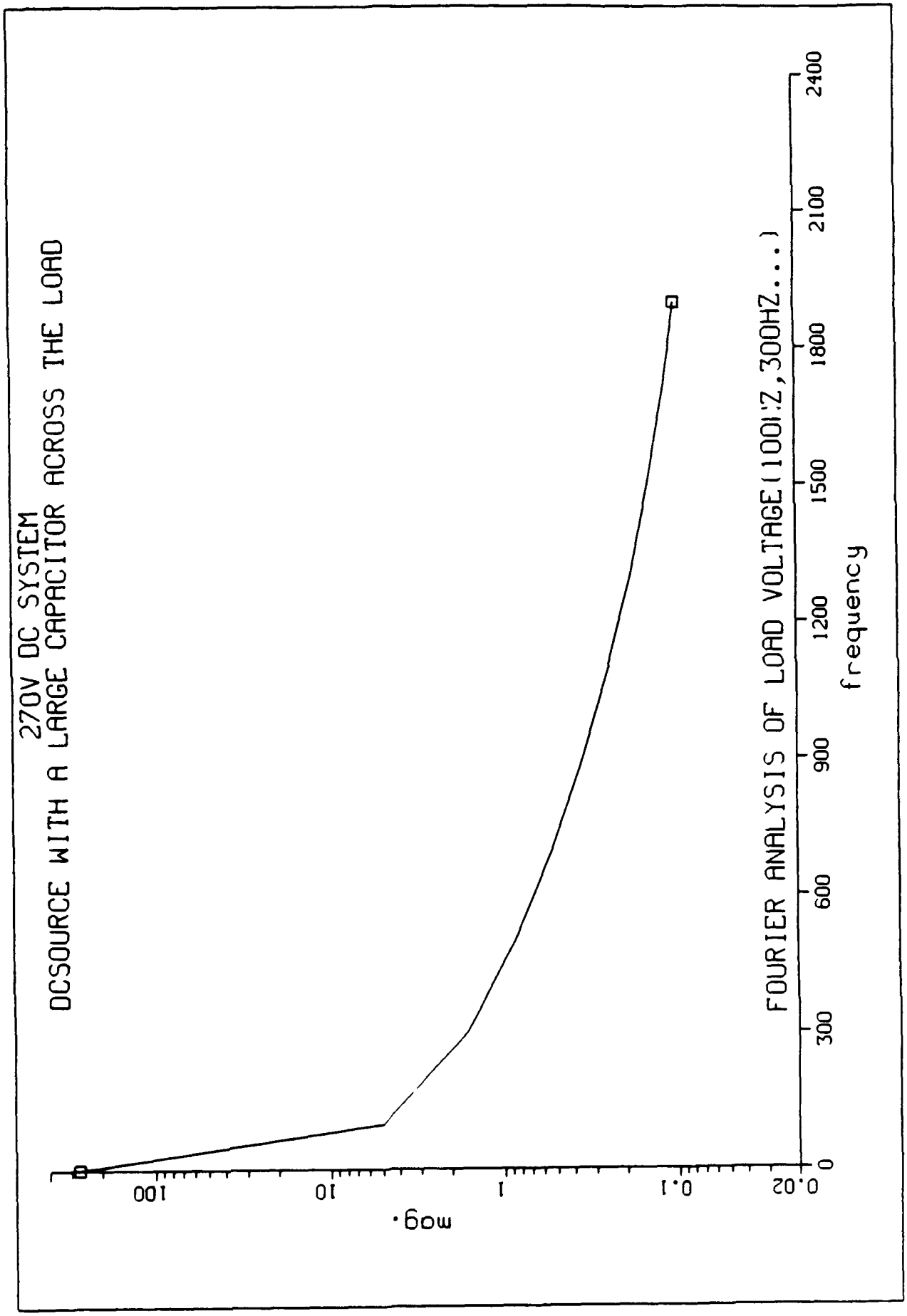

270V DC SYSTEM
DC SOURCE WITHOUT LARGE CAP OR HARMONIC FILTERS
DC 56.05
RMS 68.01
RMS(RIPPLE) 38.51
amperes
0
20
40
60
80
100
120
140
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200
seconds
CURRENT DRAWN FROM THE DC SUPPLY

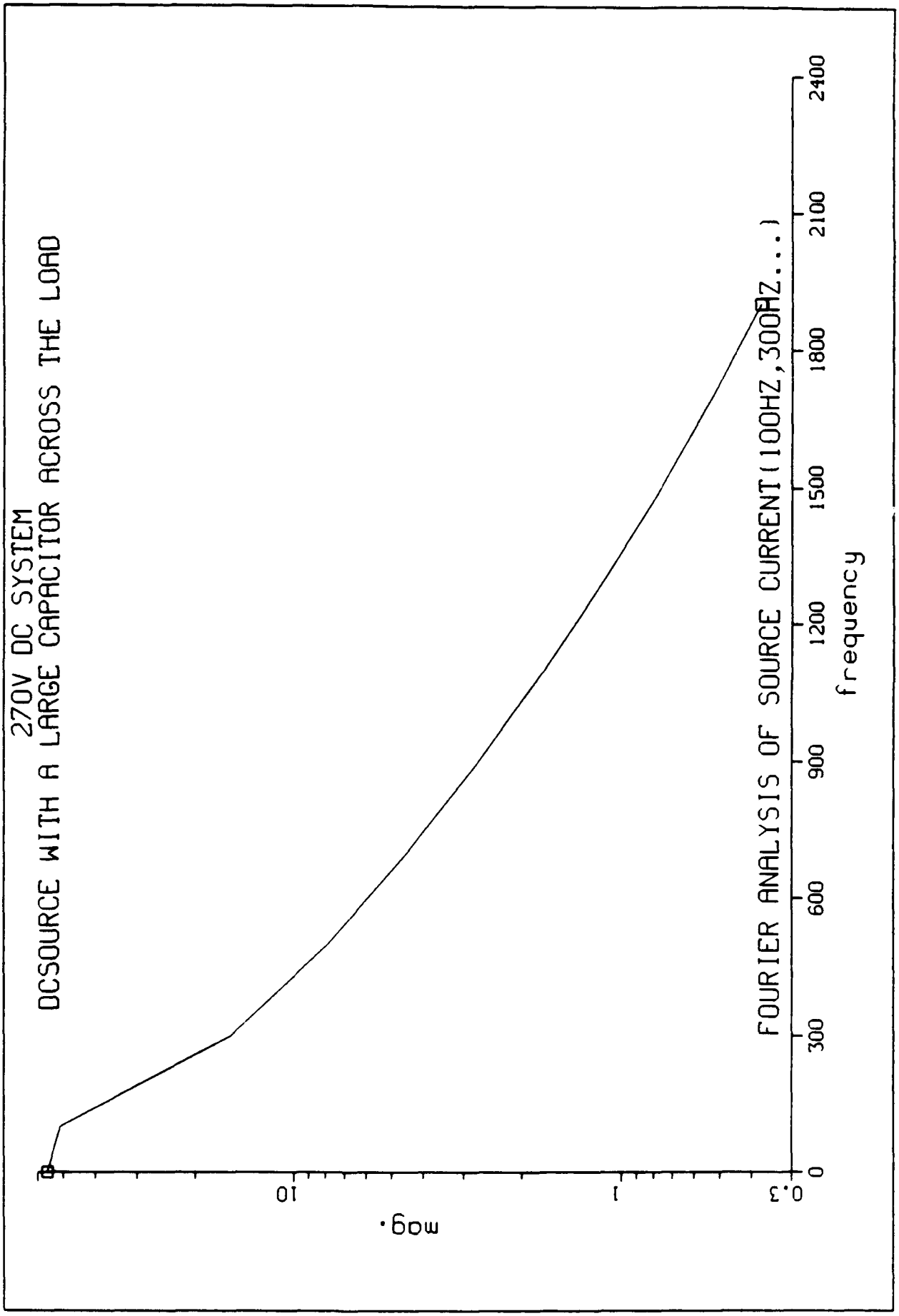
Fig. 2b

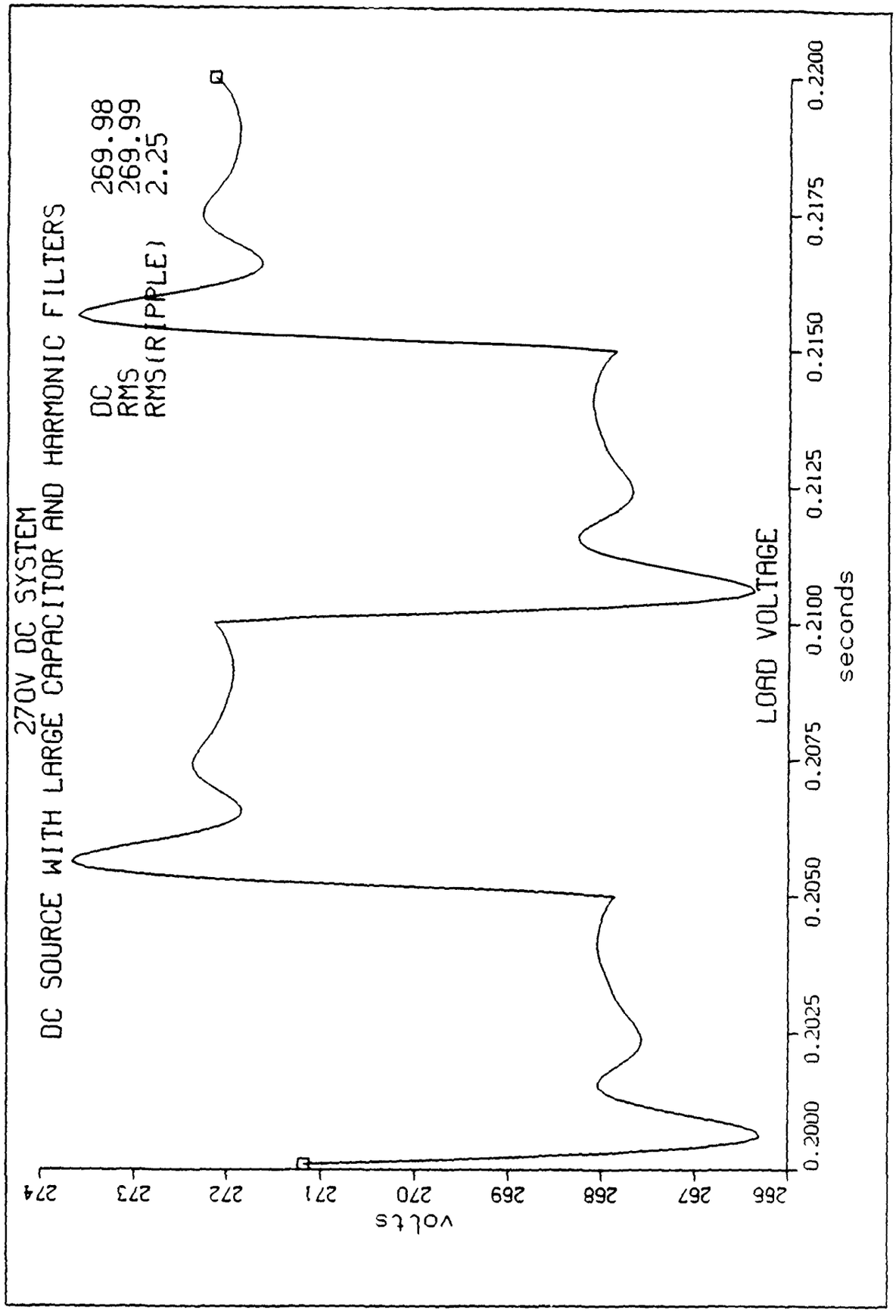

270V DC SYSTEM
DC SOURCE WITHOUT LARGE CAP OR HARMONIC FILTERS
mag.
100
10
1
0.3
0
300
600
900
1200
1500
1800
2100
2400
frequency
FOURIER ANALYSIS OF THE LOAD VOLTAGE(100HZ,300HZ...)

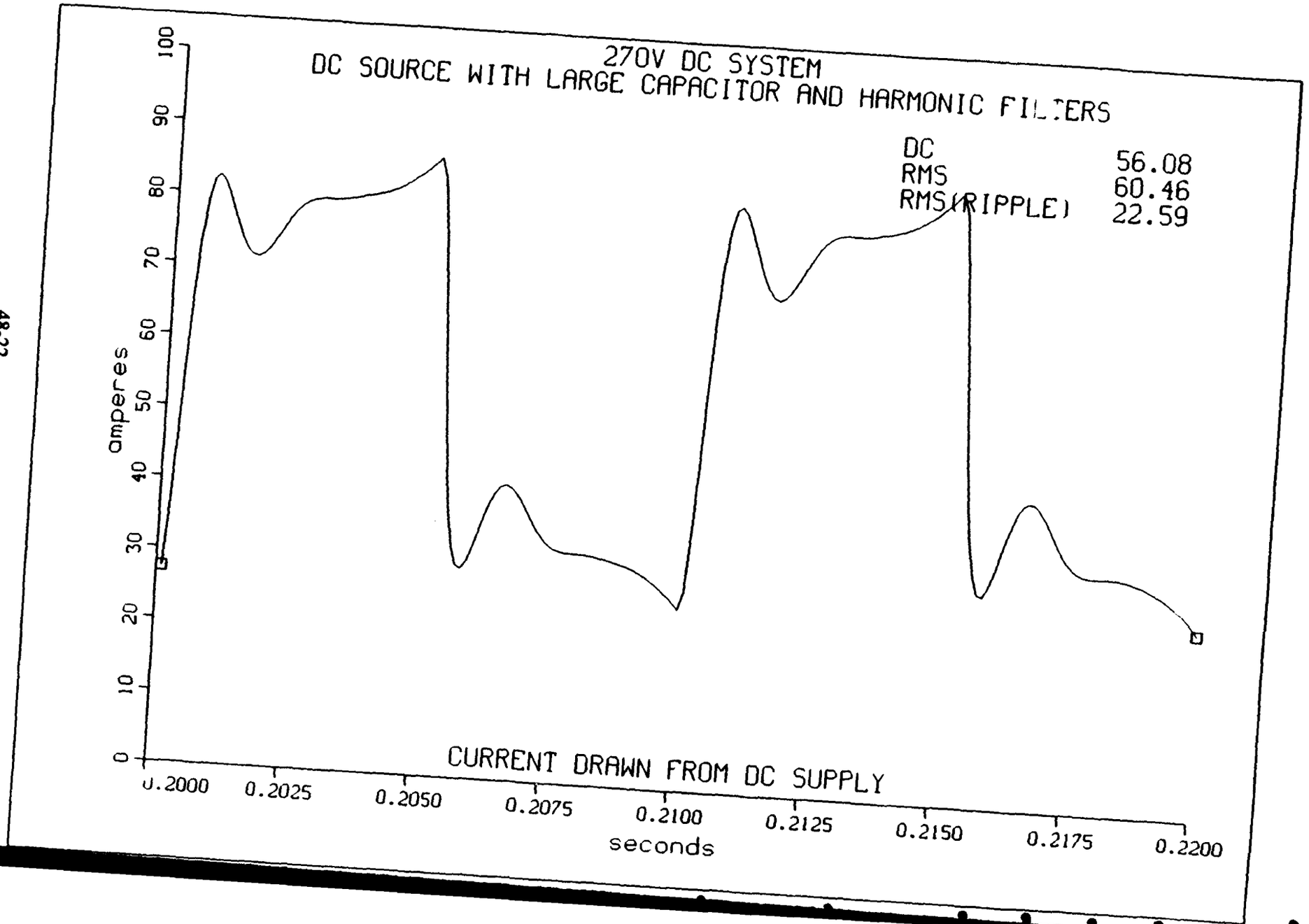

270V DC SYSTEM
DCSOURCE WITH A LARGE CAPACITOR ACROSS THE LOAD
DC 270.00
RMS 270.03
RMS(RIPPLE) 3.77
volts
276
274
272
270
268
266
264
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200
seconds
LOAD VOLTAGE

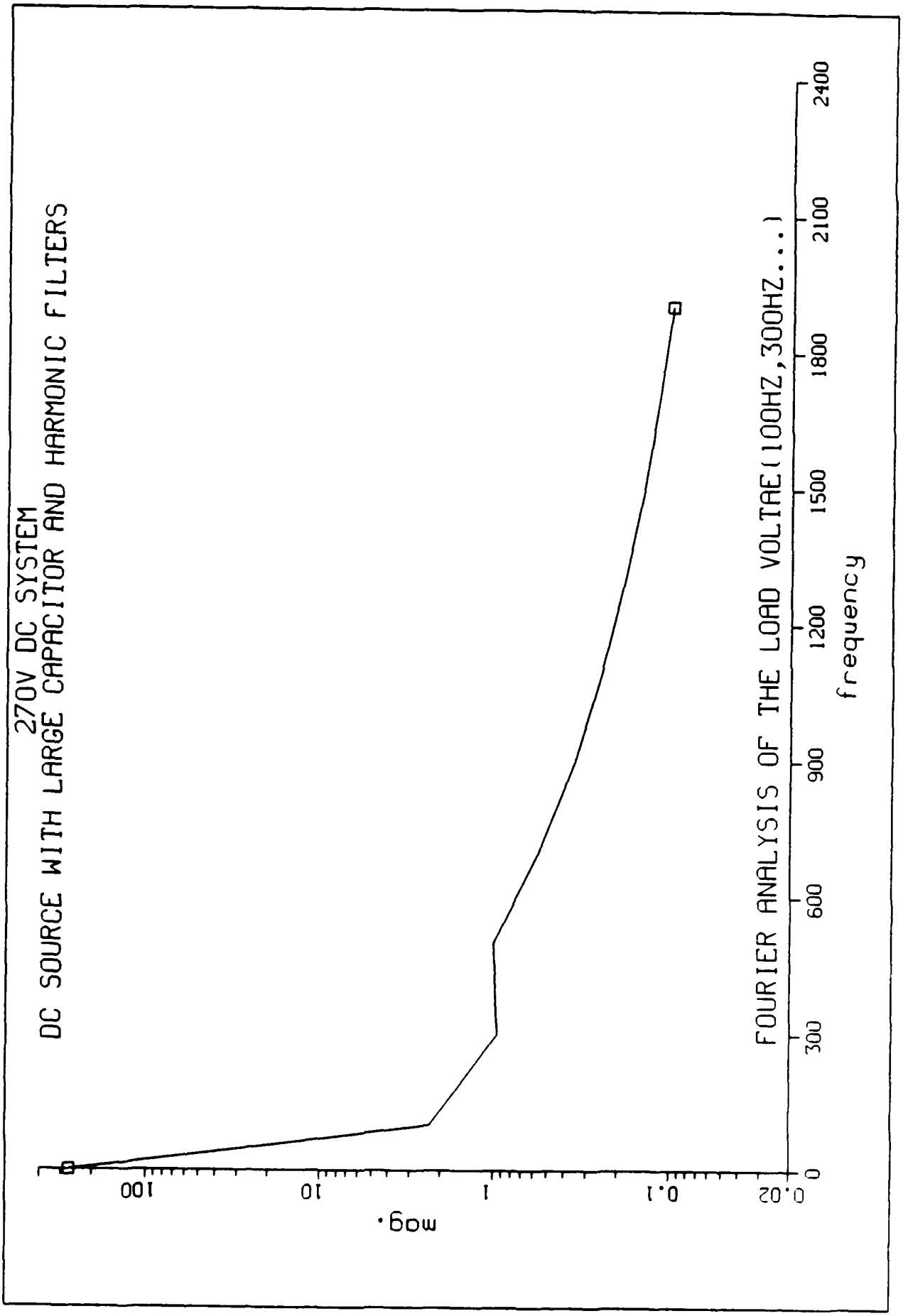

270V DC SYSTEM
DCSOURCE WITH A LARGE CAPACITOR ACROSS THE LOAD
DC 56.04
RMS 67.89
RMS(RIPPLE) 38.32
amperes
0
10
20
30
40
50
60
70
80
90
100
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200
seconds
CURRENT DRAWN FROM DC SUPPLY

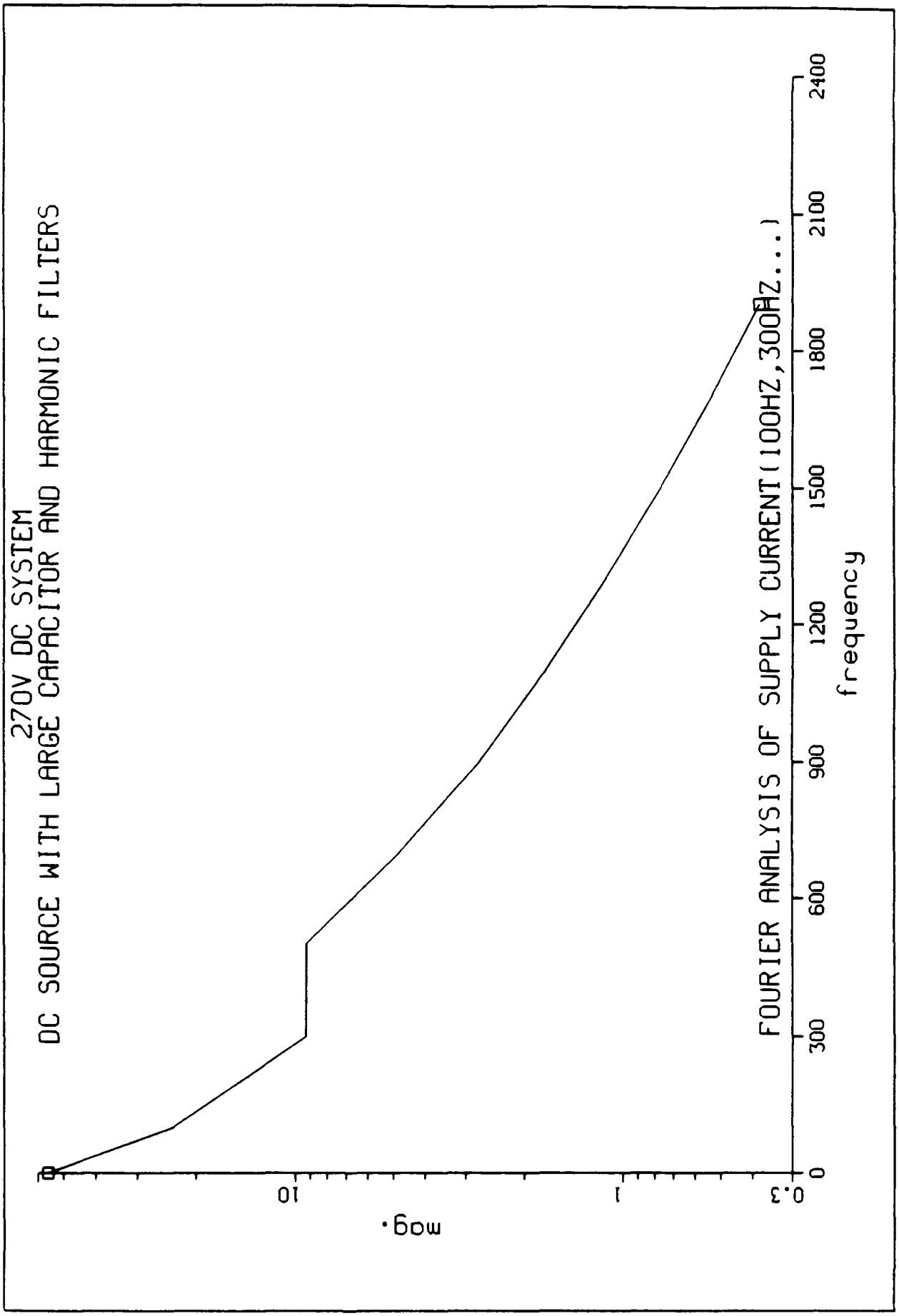

270V DC SYSTEM
DCSOURCE WITH A LARGE CAPACITOR ACROSS THE LOAD
FOURIER ANALYSIS OF LOAD VOLTAGE(100HZ,300HZ...)
mag.
100
10
1
0.1
0.02
0
300
600
900
1200
1500
1800
2100
2400
frequency

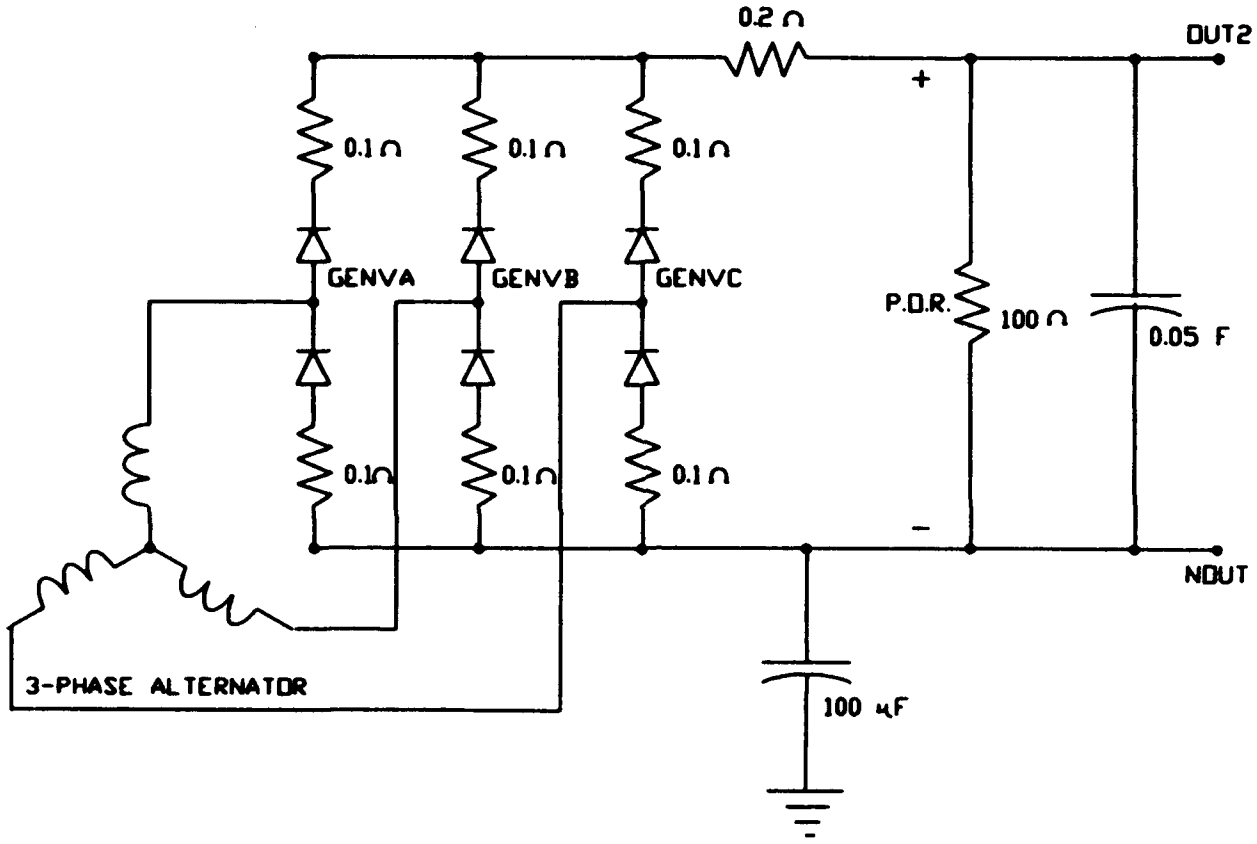

270V DC SYSTEM
DCSOURCE WITH A LARGE CAPACITOR ACROSS THE LOAD
FOURIER ANALYSIS OF SOURCE CURRENT(100HZ,300HZ...)
mag.
0.3
1
10
0
300
600
900
1200
1500
1800
2100
2400
frequency


270V DC SYSTEM
DC SOURCE WITH LARGE CAPACITOR AND HARMONIC FILTERS
DC 269.98
RMS 269.99
RMS(RIPPLE) 2.25
LOAD VOLTAGE
volts
seconds
274
273
272
271
270
269
268
267
266
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200


270V DC SYSTEM
DC SOURCE WITH LARGE CAPACITOR AND HARMONIC FILTERS
DC 56.08
RMS 60.46
RMS(RIPPLE) 22.59
Amperes
0
10
20
30
40
50
60
70
80
90
100
0.2000
0.2025
0.2050
0.2075
0.2100
0.2125
0.2150
0.2175
0.2200
seconds
CURRENT DRAWN FROM DC SUPPLY

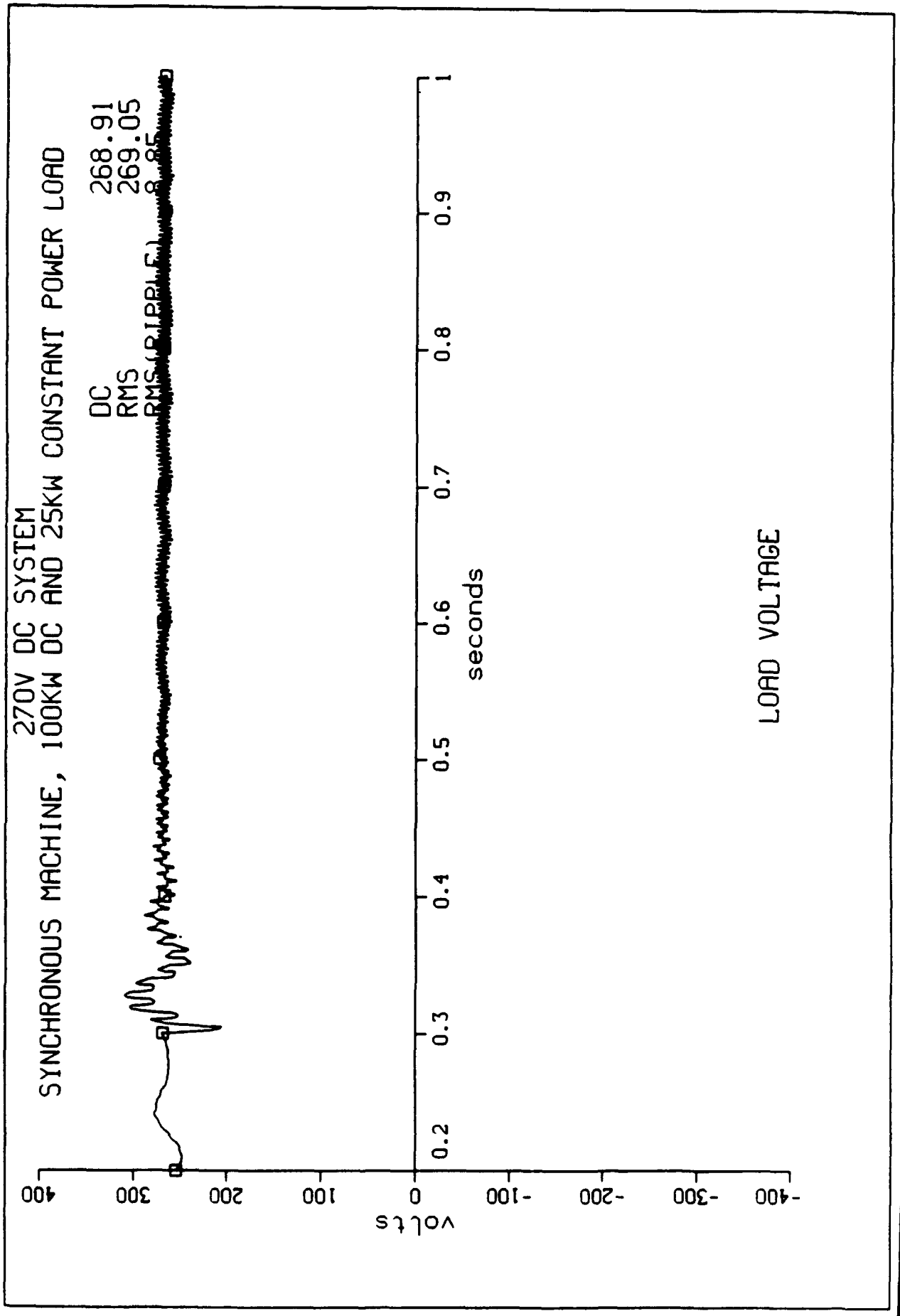

270V DC SYSTEM
DC SOURCE WITH LARGE CAPACITOR AND HARMONIC FILTERS
FOURIER ANALYSIS OF THE LOAD VOLTAE(100HZ,300HZ...)
mag.
100
10
1
0.1
0.02
0
300
600
900
1200
1500
1800
2100
2400
frequency

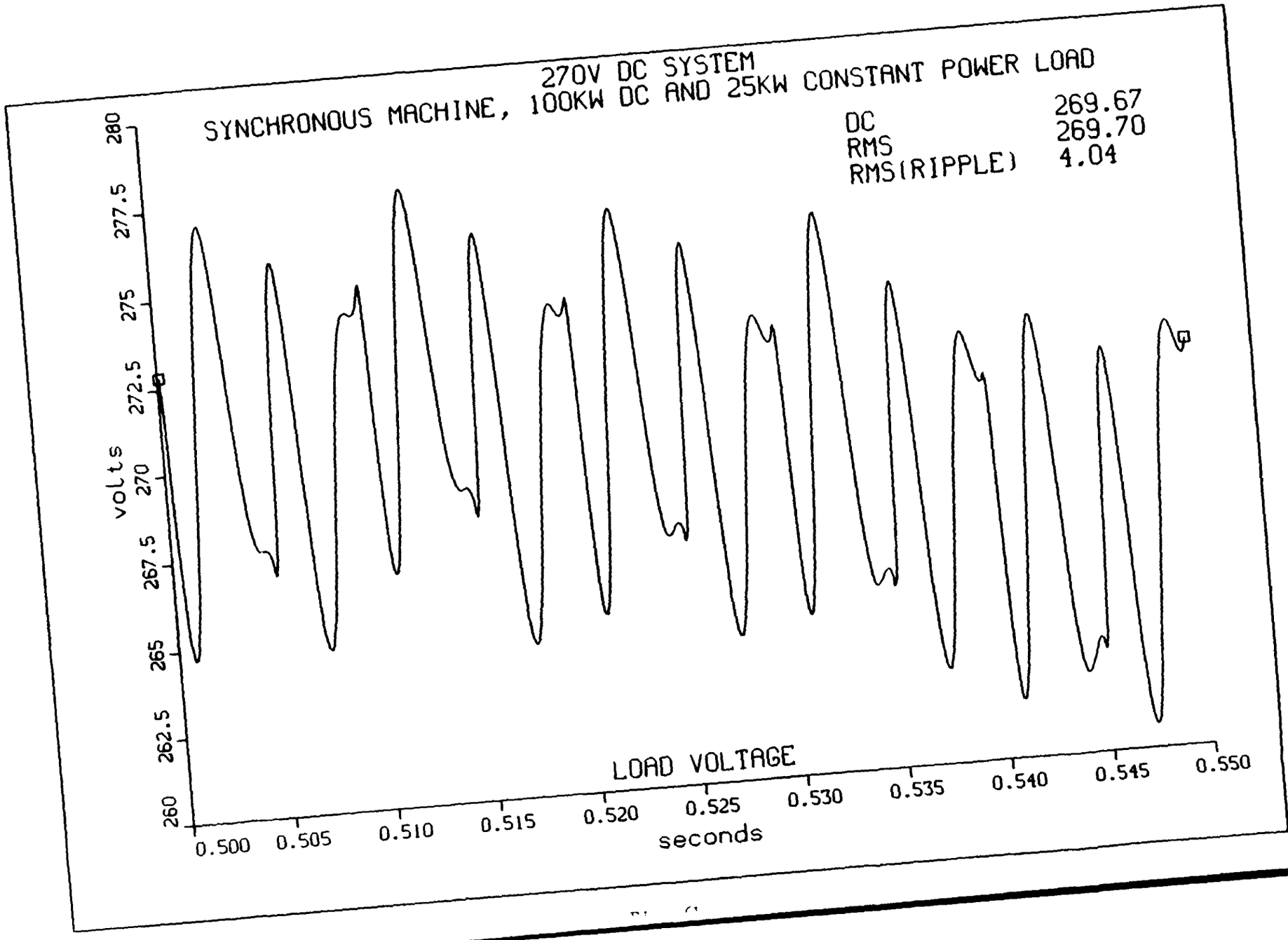

270V DC SYSTEM
DC SOURCE WITH LARGE CAPACITOR AND HARMONIC FILTERS
FOURIER ANALYSIS OF SUPPLY CURRENT(100HZ,300HZ...)
mag.
10
1
0.3
0
300
600
900
1200
1500
1800
2100
2400
frequency

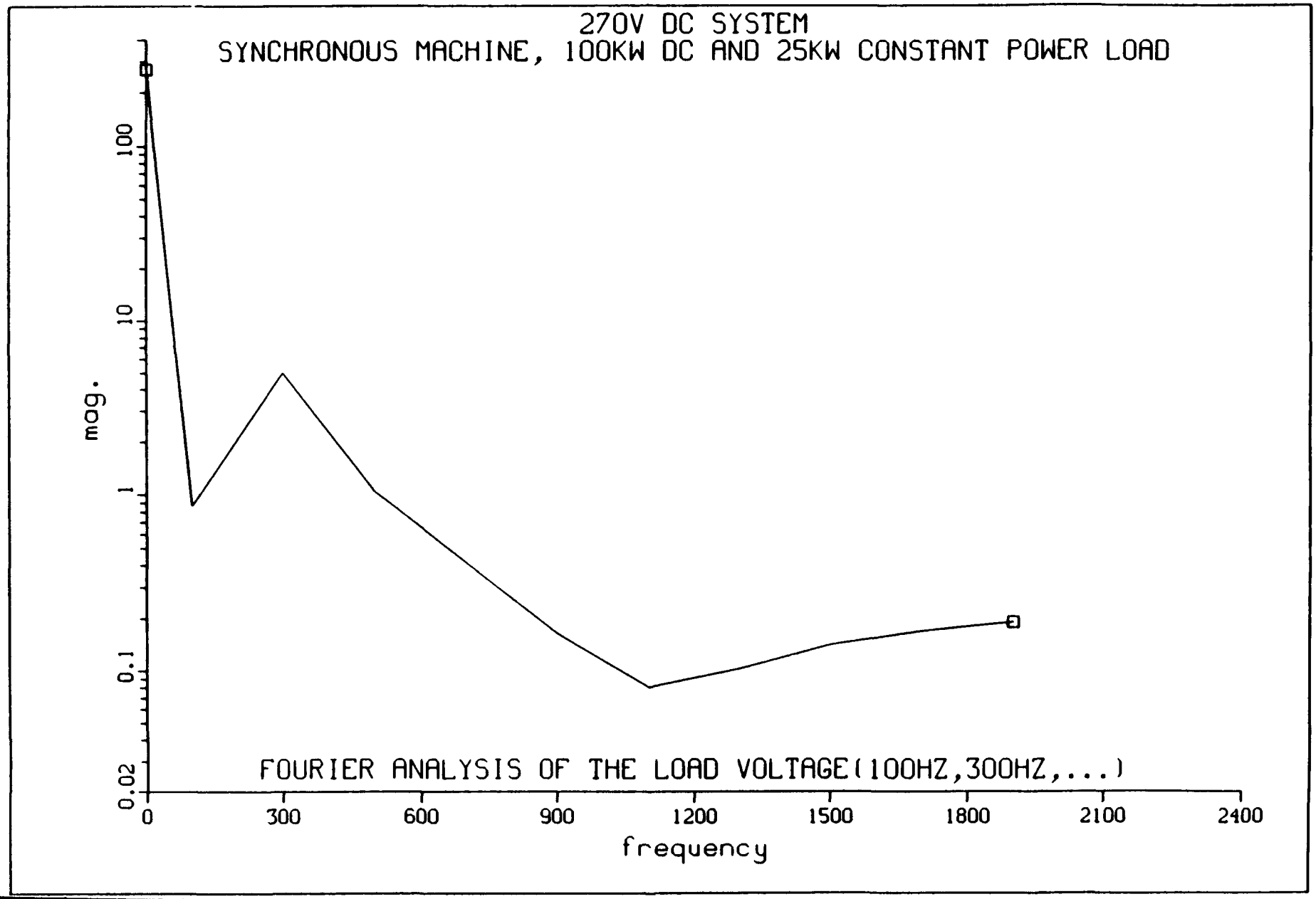


270V DC POWER SUPPLY

Fig. 5


SYNCHRONOUS MACHINE, 270V DC SYSTEM
100KW DC AND 25KW CONSTANT POWER LOAD
DC 268.91
RMS 269.05
RMS (RIPPLE)
400
300
200
100
0
-100
-200
-300
-400
volts
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
seconds
LOAD VOLTAGE


270V DC SYSTEM
SYNCHRONOUS MACHINE, 100KW DC AND 25KW CONSTANT POWER LOAD
DC 269.67
RMS 269.70
RMS(RIPPLE) 4.04
volts
280
277.5
275
272.5
270
267.5
265
262.5
260
LOAD VOLTAGE
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550
seconds


270V DC SYSTEM
SYNCHRONOUS MACHINE, 100KW DC AND 25KW CONSTANT POWER LOAD
mag.
100
10
1
0.1
0.02
0
300
600
900
1200
1500
1800
2100
2400
frequency
FOURIER ANALYSIS OF THE LOAD VOLTAGE(100HZ,300HZ,...)

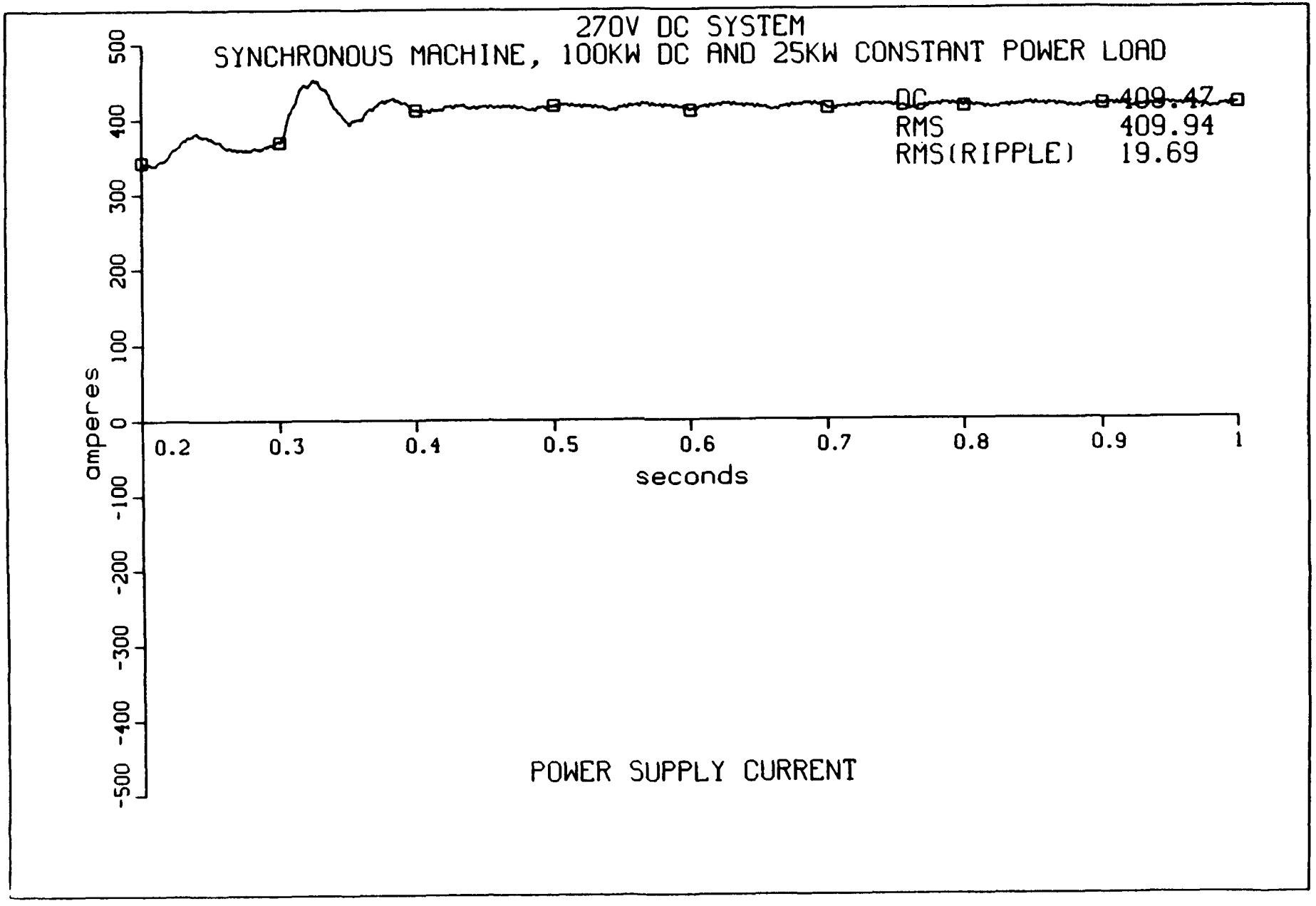
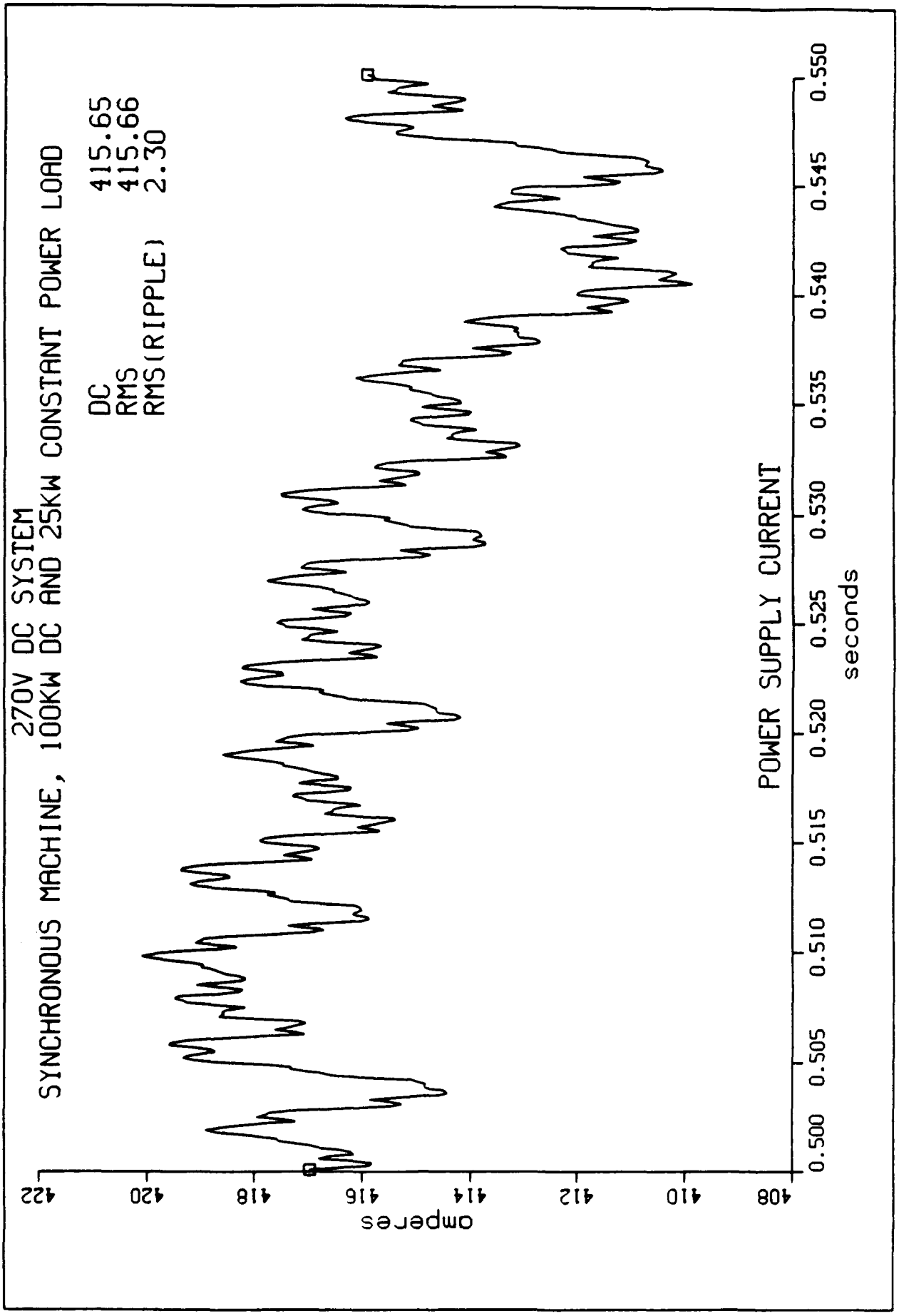
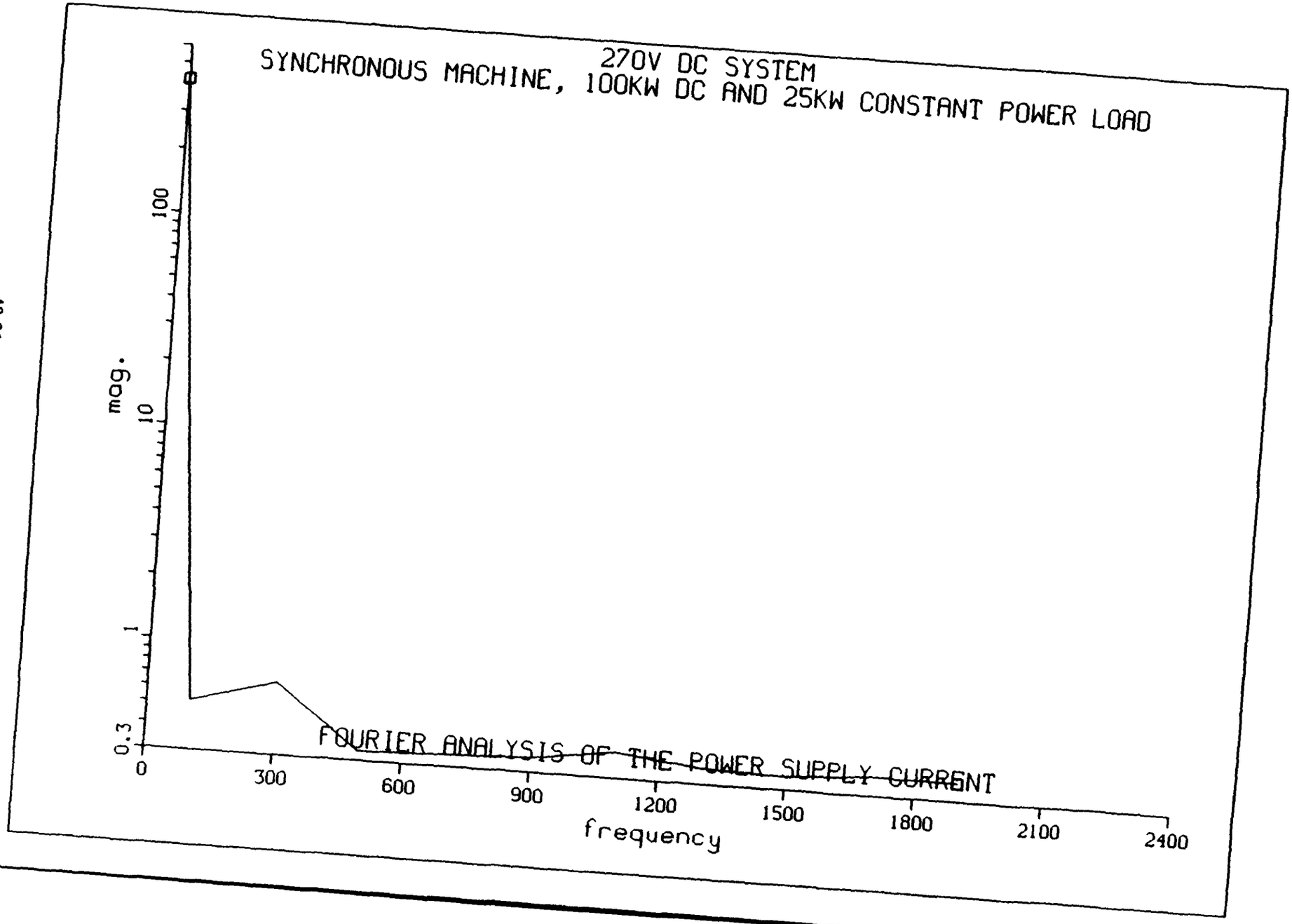
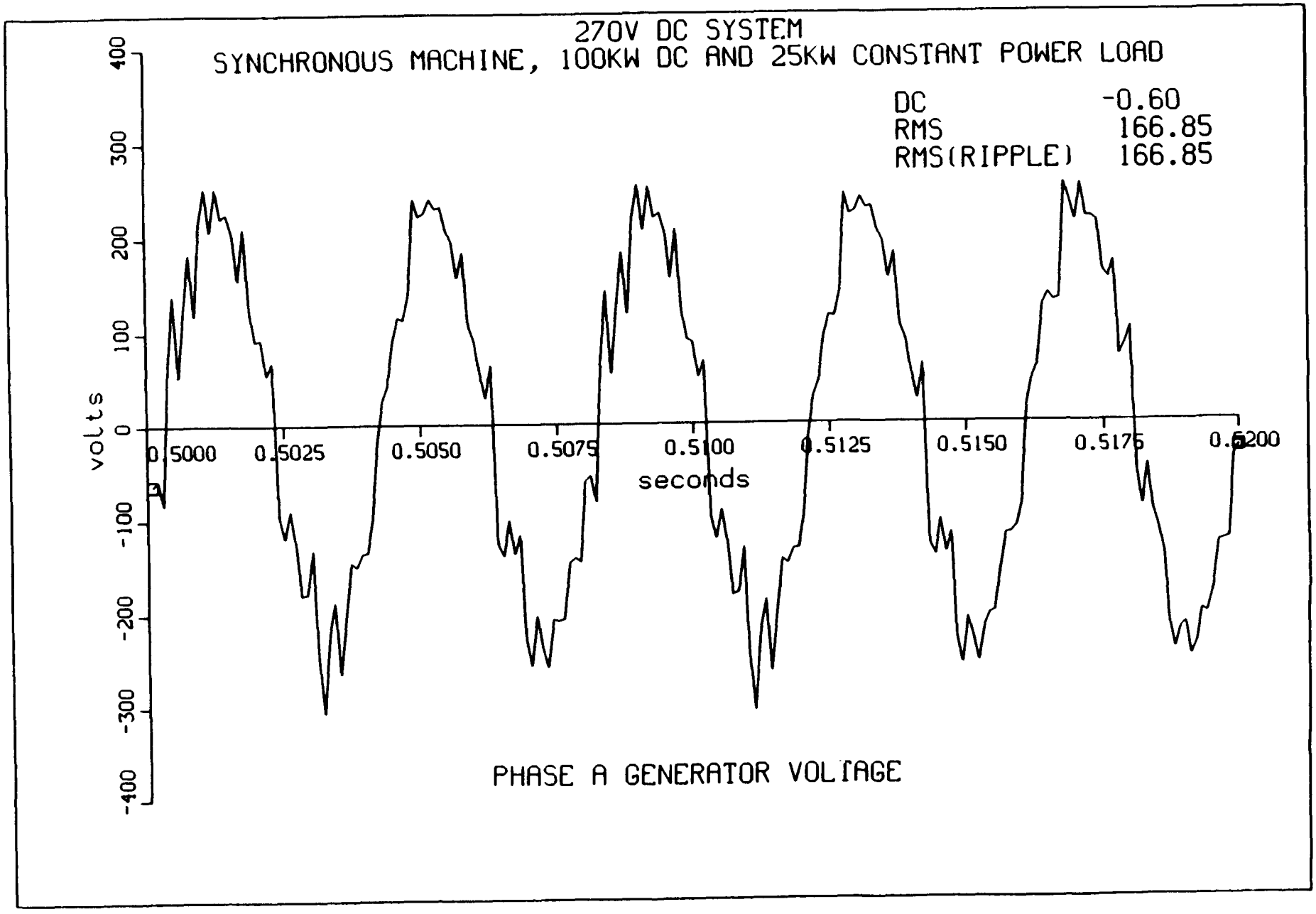

270V DC SYSTEM
SYNCHRONOUS MACHINE, 100KW DC AND 25KW CONSTANT POWER LOAD
DC 409.47
RMS 409.94
RMS(RIPPLE) 19.69
amperes
500
400
300
200
100
0
-100
-200
-300
-400
-500
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
seconds
POWER SUPPLY CURRENT

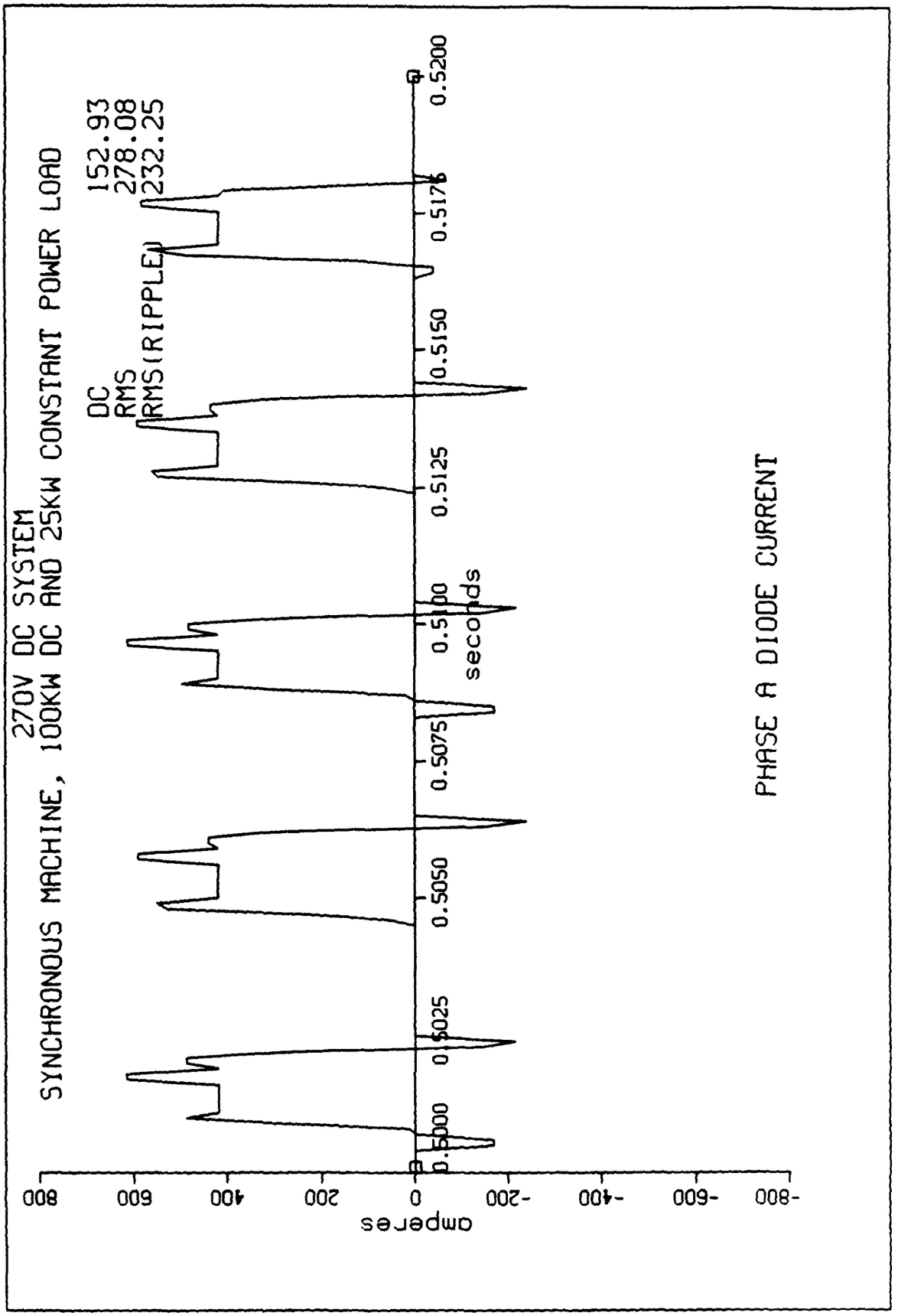
Fig. 6d

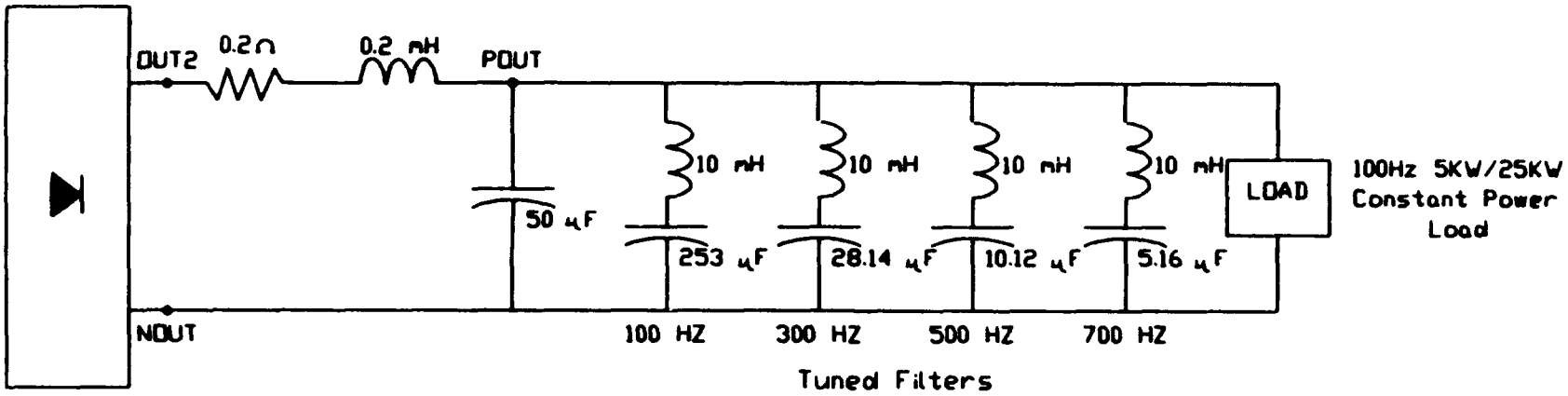

SYNCHRONOUS MACHINE, 270V DC SYSTEM
100KW DC AND 25KW CONSTANT POWER LOAD
DC 415.65
RMS 415.66
RMS(RIPPLE) 2.30
POWER SUPPLY CURRENT
amperes
seconds
422
420
418
416
414
412
410
408
0.500
0.505
0.510
0.515
0.520
0.525
0.530
0.535
0.540
0.545
0.550



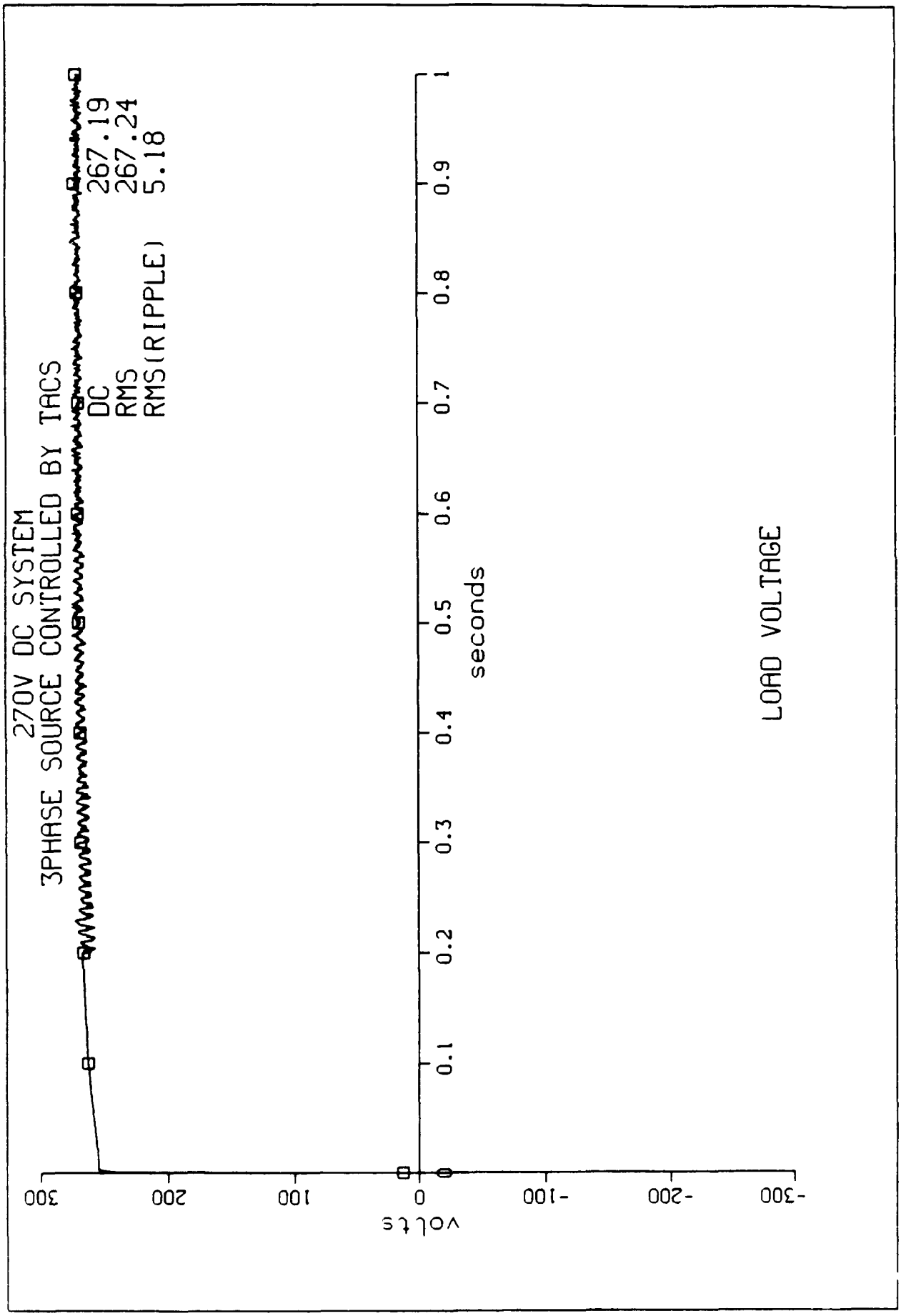
FOURIER ANALYSIS OF THE POWER SUPPLY CURRENT

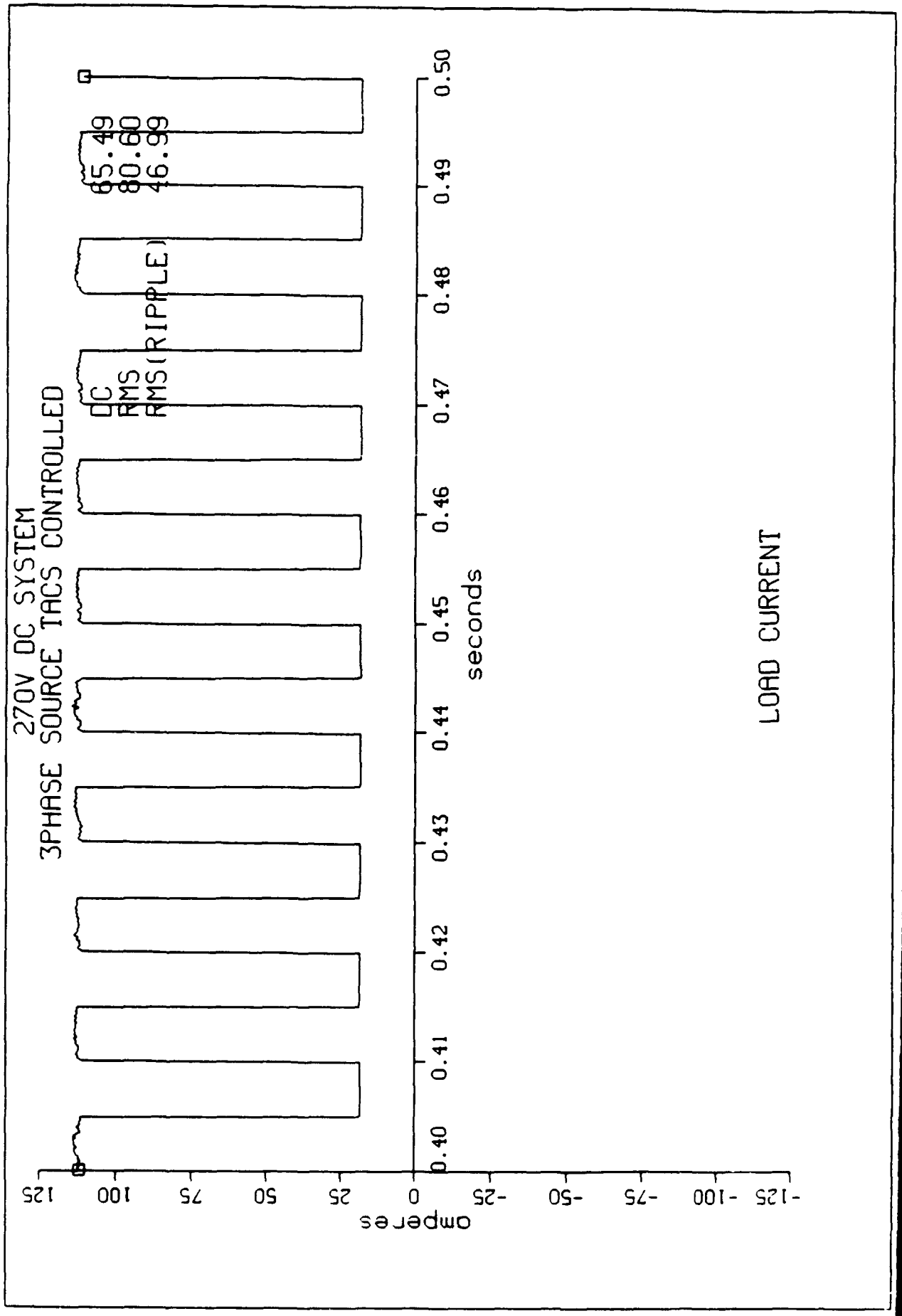

270V DC SYSTEM
SYNCHRONOUS MACHINE, 100KW DC AND 25KW CONSTANT POWER LOAD
DC -0.60
RMS 166.85
RMS(RIPPLE) 166.85
volts
400
300
200
100
0
-100
-200
-300
-400
0.5000
0.5025
0.5050
0.5075
0.5100
0.5125
0.5150
0.5175
0.5200
seconds
PHASE A GENERATOR VOLTAGE

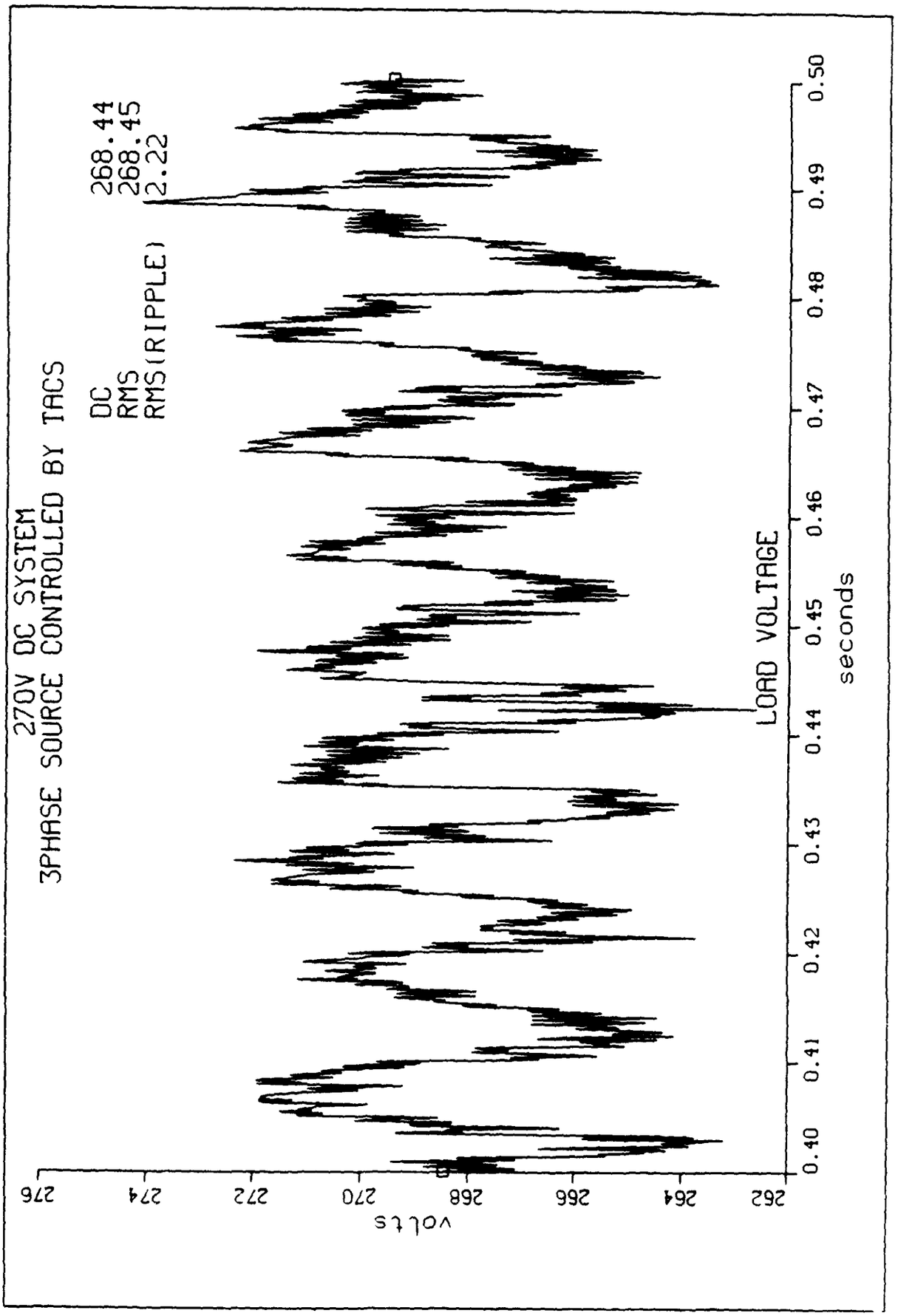

SYNCHRONOUS MACHINE, 100KW DC AND 25KW CONSTANT POWER LOAD
270V DC SYSTEM
DC 152.93
RMS 278.08
RMS(RIPPLE) 232.25
PHASE A DIODE CURRENT
amperes
seconds
800
600
400
200
0
-200
-400
-600
-800
0.5000
0.5025
0.5050
0.5075
0.5100
0.5125
0.5150
0.5175
0.5200



CONSTANT POWER LOAD WITH HARMONIC FILTERS


270V DC SYSTEM
3PHASE SOURCE CONTROLLED BY TACS
DC 267.19
RMS 267.24
RMS(RIPPLE) 5.18
LOAD VOLTAGE
seconds
volts
300
200
100
0
-100
-200
-300
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


270V DC SYSTEM
3PHASE SOURCE TACS CONTROLLED
DC 65.49
RMS 80.60
RMS(RIPPLE) 46.99
seconds
0.40
0.41
0.42
0.43
0.44
0.45
0.46
0.47
0.48
0.49
0.50
amperes
125
100
75
50
25
0
-25
-50
-75
-100
-125
LOAD CURRENT


270V DC SYSTEM
3PHASE SOURCE CONTROLLED BY TACS
DC 268.44
RMS 268.45
RMS(RIPPLE) 2.22
LOAD VOLTAGE
volts
seconds
262
264
266
268
270
272
274
276
0.40
0.41
0.42
0.43
0.44
0.45
0.46
0.47
0.48
0.49
0.50

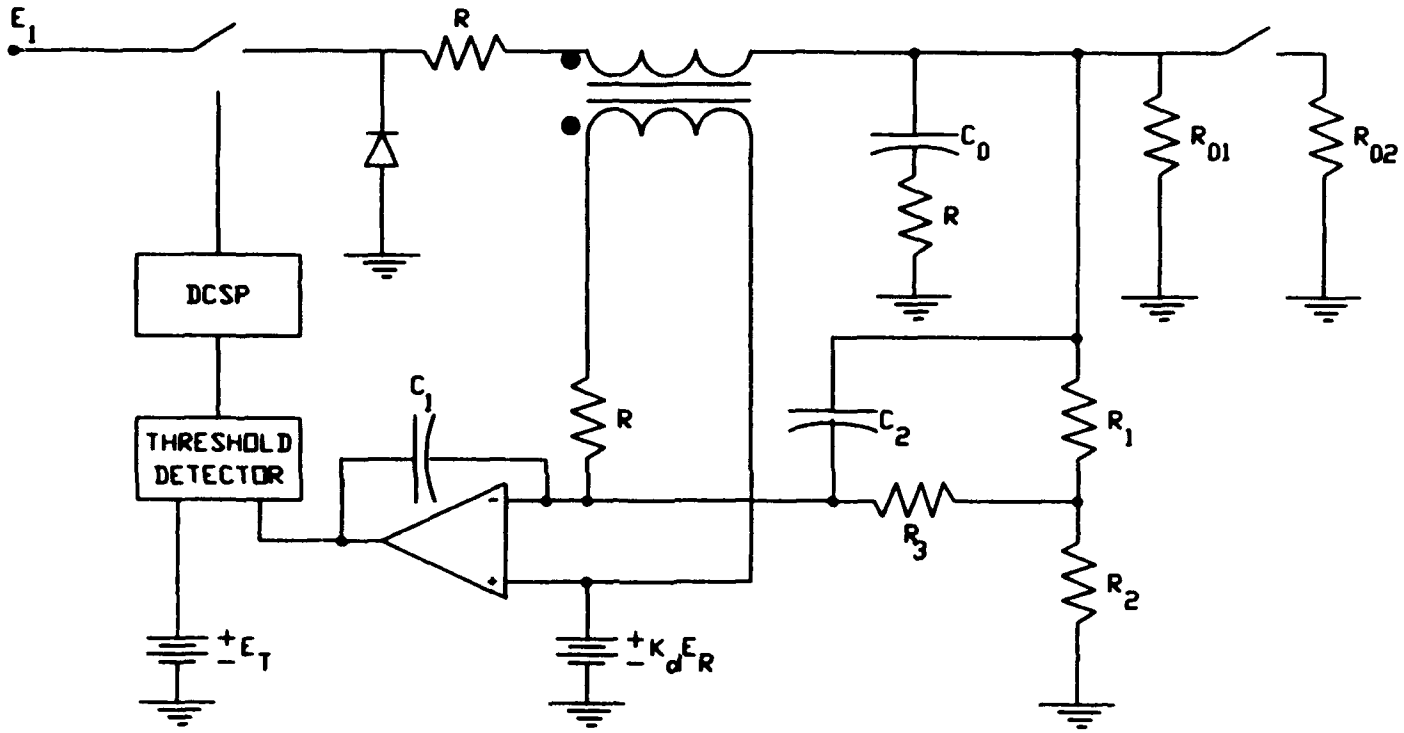

E1
R
DCSP
THRESHOLD DETECTOR
C1
R
C0
R
R01
R02
C2
R1
R3
R2
+ ET -
+ KdER -

Fig. 10

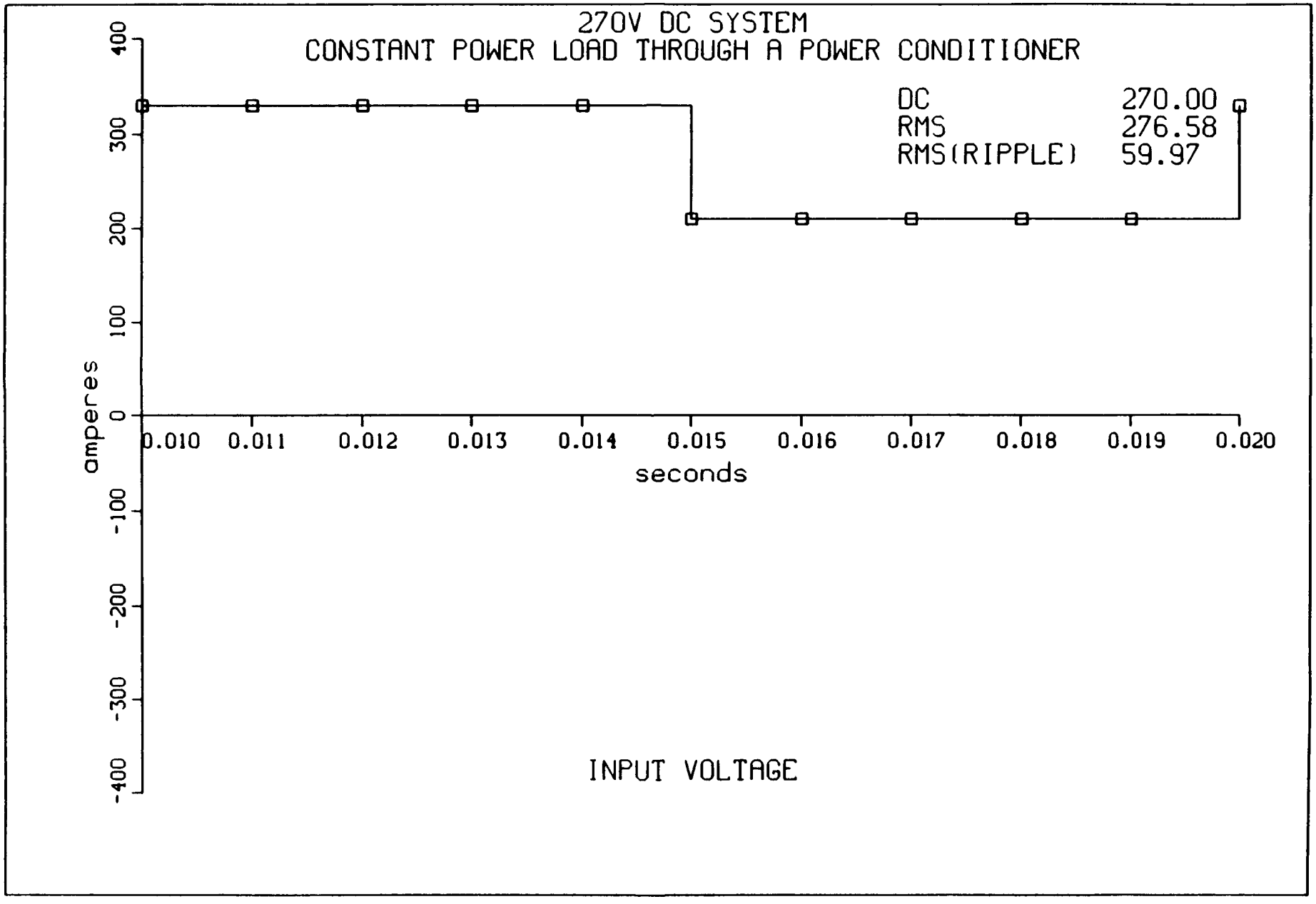
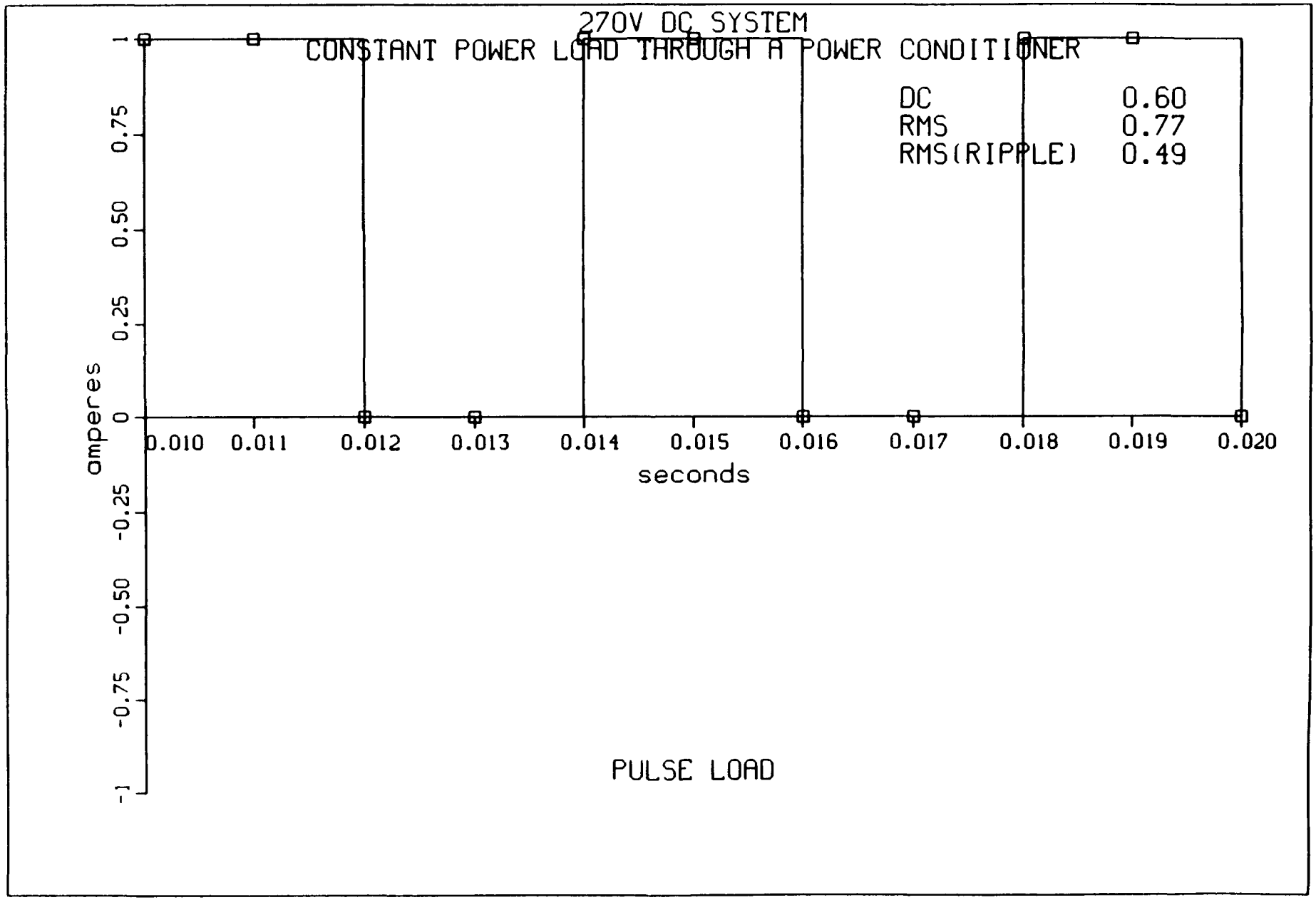

270V DC SYSTEM
CONSTANT POWER LOAD THROUGH A POWER CONDITIONER
DC 270.00
RMS 276.58
RMS(RIPPLE) 59.97
amperes
400
300
200
100
0
-100
-200
-300
-400
0.010
0.011
0.012
0.013
0.014
0.015
0.016
0.017
0.018
0.019
0.020
seconds
INPUT VOLTAGE

Fig. 11a


270V DC SYSTEM
CONSTANT POWER LOAD THROUGH A POWER CONDITIONER
DC 0.60
RMS 0.77
RMS(RIPPLE) 0.49
amperes
1
0.75
0.50
0.25
0
-0.25
-0.50
-0.75
-1
0.010
0.011
0.012
0.013
0.014
0.015
0.016
0.017
0.018
0.019
0.020
seconds
PULSE LOAD

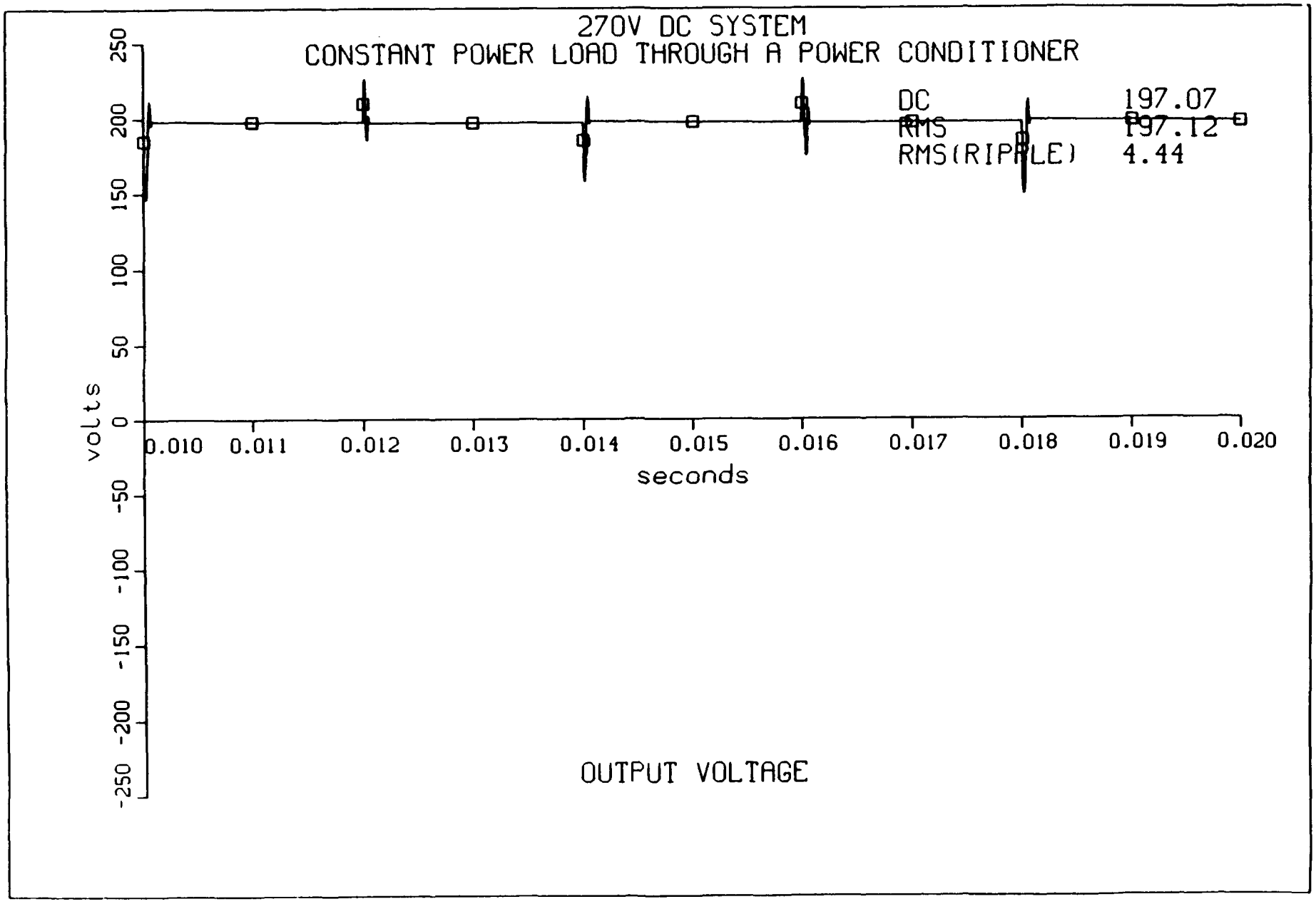
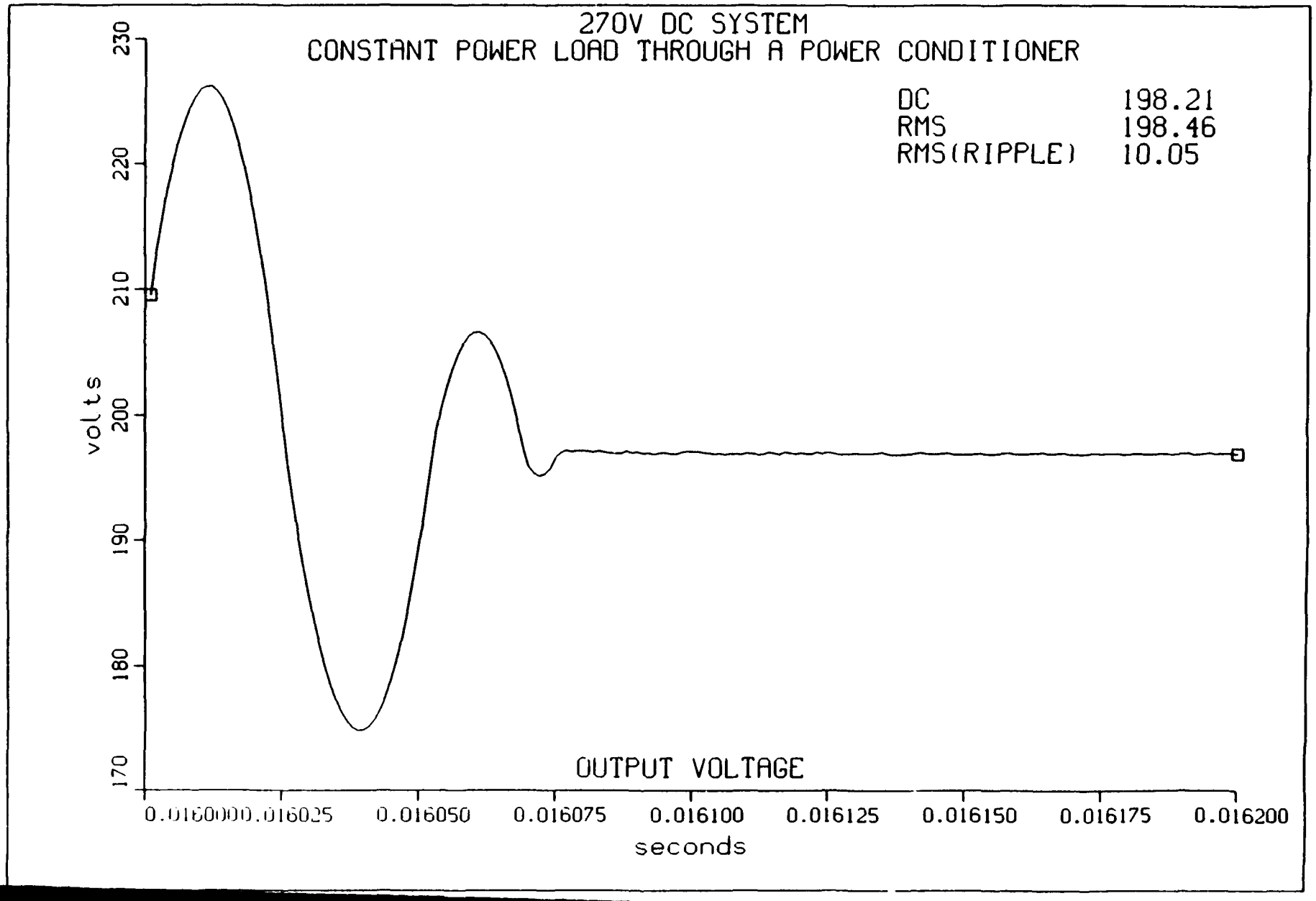

270V DC SYSTEM
CONSTANT POWER LOAD THROUGH A POWER CONDITIONER
DC 197.07
RMS 197.12
RMS(RIPPLE) 4.44
volts
250
200
150
100
50
0
-50
-100
-150
-200
-250
0.010
0.011
0.012
0.013
0.014
0.015
0.016
0.017
0.018
0.019
0.020
seconds
OUTPUT VOLTAGE

Fig. 11.


270V DC SYSTEM
CONSTANT POWER LOAD THROUGH A POWER CONDITIONER
DC 198.21
RMS 198.46
RMS(RIPPLE) 10.05
OUTPUT VOLTAGE
volts
230
220
210
200
190
180
170
0.016000
0.016025
0.016050
0.016075
0.016100
0.016125
0.016150
0.016175
0.016200
seconds

DYNAMIC TEMPERATURE MEASUREMENTS IN REACTING FLOWS

UES Contract F49620-88-C-0053/SB5881-0378

Purchase Order No. S-210-11MG-001

Final Report for Contract Period 1/1/91 - 12/31/91

Mechanical Engineering Department

Virginia Polytechnic Institute and State University

Blacksburg, Virginia

Thomas R. Scattergood, Graduate Research Assistant

Larry A. Roe, Principal Investigator

INTRODUCTION

When studying combustion systems and reacting flows, it is important to be able to measure spatially and temporally resolved temperatures. This data, along with corresponding velocity data, aides in the understanding of various aspects of fluid flow such as stability characterization and control, interpretation of velocity data, and the understanding of turbulence correlations and compressibility effects. Since these fluid phenomena are of a rapidly changing nature, it is important to have a temperature measurement rate sufficient to keep up with such fluctuations in the flowfield. In this case, a thermal data rate of about 1 kHz would be required for the anticipated application to research ramjet combustors.

Current optical techniques based on the quantum-mechanical effects of light are useful in their ability to provide temperature measurements in flows due to their non-intrusiveness, their accuracy, and their speed. They can, however, be prohibitively expensive and very complicated, and none can yet provide combined spatial and temporal resolution at a satisfactory sampling rate. Thermocouples, on the other hand, are relatively inexpensive and uncomplicated and a great deal of past experience provides a base of knowledge upon which to rely. There still remain problems involving probe survivability and disturbance, as well as the usual worries over data validity, but the current research is attempting to provide insight into these areas of concern.

Thermocouples are able to provide very accurate information about the temperature of the probe junction, but there is a complex relationship between the junction temperature and the instantaneous gas temperature. Since the physical size of any thermocouple prevents it from responding directly to temperature fluctuations at frequencies above about 200 Hz, compensation is necessary. For the case of minimal radiation and conduction losses, it may be shown that:

$$Tc = Tj + \alpha(dTj/dt)$$

where Tc is the actual gas temperature, Tj is the measured junction temperature, and α is referred to as the time constant, which represents a ratio between the junction thermal capacity and the convective heat flux. Since α is a function of the instantaneous gas properties, calculating the time constant directly would not normally be possible, and most researchers have used some sort of average value. This has proven to be insufficient and an alternate method of determining α, and thus the instantaneous gas temperature, is needed.

One possible solution is to use a two-wire thermocouple probe in which two junctions of differing diameters are placed close together. This technique, theoretically, allows the determination of an instantaneous time constant based on the temporal temperature gradients of each junction. The focus of this research is to prove this technique and apply it to an actual flow with possible future correlation to a CARS system and then to an LDV for simultaneous velocity measurements.

LITERATURE REVIEW

An extensive review of thermocouple techniques has been conducted and an abbreviated version of this review, limited to those studies relevant to the proposed program, is presented. The primary items of interest in this review are: flowfield type, size of junction or wire, coating, radiation and conduction corrections, temporal resolution, and response compensation. Virtually all the referenced experimental studies used wires of Pt and/or Pt-Rh.

Fristrom and Westenberg (1965) provide detailed instructions on the fabrication of small thermocouples and generating silica coatings for the elimination of catalysis effects. The coating technique involves passing the junction through a flame containing silicon dioxide generated from the combustion of a compound such as dimethyl siloxane. Thermocouples prepared using this method were evalutated in hydrogen-air flames by Cookson et al. (1964), with significant errors in the mean measured temperature for the larger

thermocouples attributed to recombination reactions in the boundary layer. Radiation corrections were implemented, but transient response effects were not considered. Observed differences between coated and uncoated thermocouples due to surface catalysis were on the order of 400C, for mean temperatures of 1100C.

Kent (1970) reported that silica is not appropriate for temperatures above about 1100C, as free silicon diffuses into the junction, altering the properties of the thermocouple. A yttria-beryllia coating was suggested as an alternate, tested, and found to give good results.

Yule et al. (1978) offer a good discussion of transient compensation concepts, and provide a correction for non-cylindrical junctions. (Heat transfer to a cylinder has often been assumed in predicting thermal response.) Average (not instantaneous) time constants for the actual junctions were determined by observing the response to a step-change in temperature by measuring the decay in measured temperature after a heating current was switched off. It was stated that this technique will not work unless the thermocouple wires to the junction are much smaller than the probe lead wires. Wire size and junction size were 25 and 70 microns, respectively; a typical time constant was 30 msec. Some reasonable agreement was achieved between predicted average time constants based on Nusselt number relations and measured data. In a propane-air diffusion flame, average time constants were determined at different locations in the flowfield by pulsed overheating (again, a response to a known step change in temperature) and used to correct the observed temperatures. These thermocouples were apparently not coated. Independent confirmation of the observed temperatures was not provided. A two-wire technique, with one wire constantly heated to provide heat transfer data for response correction, was suggested as a future development.

Lockwood and Odidi (1973) measured mean temperatures in a turbulent diffusion methane flame with a 40 micron thermocouple. No corrections for radiation or conduction were made and the thermocouple was not coated. An average time constant for each location in the flame was determined by an AC excitation method, a

typical value was 10 msec. This time constant was then applied to an on-line compensation circuit to obtain a corrected voltage. The use of an average, rather than instantaneous, time constant was estimated to generate errors of up to 10 percent in the mean temperature and 20 percent in the instantaneous temperature.

Lockwood and Moneib (1980) measured the temperature in a nonreacting, electrically heated free jet with a 12.7 micron thermocouple compensated for transient response. (they also concluded that the results of the previous work were not as good as originally claimed.) An on-line, digital compensation network was developed; the appropriate time constant was still an average value determined by a pulsed overheat method. The time constant was on the order of 15 msec; noise limited the max frequency response to 5 kHz. Prodigious quantities of results were plotted, including mean profiles, PDF's, flatness, and skewness, and were described as "physically realistic."

El Banhawy et al. (1983) measured mean values of temperature, velocity, and concentrations in step-stabilized, premixed, methane-air flames. The thermocouples were 15, 40 or 80 micron diameter, with the maximum indicated temperature difference between the three being 40C. Conduction effects were claimed unimportant as a result of the probe design, radiation effects were not addressed, the junctions were apparently not coated, and no correction for response was implemented. A frequency response to 200 Hz was claimed.

Heitor et al. (1985) measured temperature simultaneously with LDV data in premixed, disc-stabilized, natural gas flames. The maximum error due to radiation was estimated to be 100K; a specific correction was not made. The thermocouple output was digitized and stored, with the compensation being done off-line by determining the gradient dTj/dt from the stored data. An average time constant was not used, rather, an instantaneous time constant was determined based on the measured velocity, physical characteristics of the junction, and a heat transfer relation for fine wires based on the results of Collis and Williams (1959), with some revisions. This was a good idea, but didn't work very well in the reacting flows, as the gas

properties are not uniquely determined by the temperature because of large variations in chemistry. Plots of temperature PDF's near the edge of the flame showed extreme overcompensation by this method, so the "instantaneous" time constants were all multiplied by 0.65 (trial-and-error value) to make the PDF's look better. The junctions were left uncoated to keep the time constants as low as possible. Significant build-up of the titanium dioxide seeding particles on the thermocouple surface caused the response to degrade rapidly with run time, with LDV data rates of 100/sec being high enough to render the compensation technique essentially useless. The wire diameters were again 15, 40, or 80 microns.

As part of a modeling verification, Abdalla et al. (1982) used silica-coated, 100 micron thermocouples to measure mean temperature in premixed, methane-air flames. Uncoated, 50 micron wires were used for temporally resolved temperatures. An average time constant was determined at each location in the flowfield by an overheat method and dialed into an electronic compensation RC network, responsive to 5 kHz. The accuracy of these results was not discussed to any significant extent.

Masri and Bilger (1984) measured average temperatures in hydrocarbon turbulence diffusion flames using an uncoated, uncompensated thermocouple with a 130 micron wire diameter and 270 micron bead diameter. Radiation errors were estimated to be 30C maximum. Masri and Bilger (1986) used this same system to determine average temperatures in a different burner. Starner and Bilger (1986) measured mean temperatures in a swirling hydrogen diffusion flame with a 200 micron (bead size) thermocouple, corrected for radiation losses. Transient compensation was again ignored. Neither the absolute accuracies nor effect of the probe on the flow was discussed.

Yoshida and Tsuji (1978) evaluated the temperature and velocity distributions in a premixed propane flame, but not simultaneously. The thermocouple was 50 micron diameter and uncoated. Transient compensation was accomplished in a conventional manner (initially) by determining the time constant from an overheat-response method, then incorporating this value in

a 5 kHz, RC compensation circuit. This gave maximum instantaneous temperatures higher than adiabatic flame temperatures, and minimum termperatures lower than ambient. To correct this discrepancy, a series of different time constants was used in the compensation circuit until a value was found which normalized the output between the ambient and adiabatic flame temperatures. This value (40 msec) was then used throughout the flowfield. Some "mismatching" was expected by the authors from this procedure. Temperature fluctuations of 400C, at frequencies above 1 kHz, were found.

Katsuki et al. (1987) developed a linearization technique for time-response compensation and evaluated the procedure by rapidly vibrating the thermocouple juntion across the flame front of a laminar diffusion flame. Both coated and uncoated wires were used. The technique utilized for compensation essentially determines a film temperature for the wire and evaluates the instantaneous time constant based on that temperature. Variations in chemistry are apparently not accounted for. Differences in measured values between coated and uncoated junctions were ascribed to radiative effects in laminar diffusion flames and catalysis in turbulent premixed flames. Results from the new technique were compared to results using a standard, average time constant compensation technique. Neither procedure was found to give especially impressive measurements; the authors concluded that simultaneous velocity data were necessary to properly account for variations in the time constant. It was also concluded that coating did not adversely affect the determination of real-time data; although the time constant was increased due to the larger diameter, compensation produced the same temperature PDF's as for the compensated, uncoated thermocouple.

Lenz and Gunther (1980) used an uncoated, 50 micron, frequency compensated thermocouple to determine time-resolved temperature in a free-jet diffusion flame. Compensation was accomplished by determining the average time constant at each position in the flow by an overheat-response method, then using the resulting value in an on-line electrical compensation network.

Conduction, radiation, and catalytic effects were not corrected. Response to 8 kHz was obtained, limited by noise. Several apparent discrepancies in the temperature data were explained away by physical arguments.

Brum et al. (1983) used a 25 micron, uncoated, compensated thermocouple to acquire temperature data simultaneously with LDV data in a swirling, reactive flow. The compensation was similar to that of Lockwood and Moneib, using an in-situ overheat method to determine an average time constant and on-line electronic compensation of the signal. The authors indicated that this average time-constant approach was not optimum, and attributed some apparent discrepancies in the data to the instrumentation. It was determined that a variation of 10 percent in the time constant could lead to errors as high as 50 percent in local heat flux. An alternative technique, utilizing the instantaneous velocity to modify the time constant, was suggested. (this approach was later used by Heitor et al. with minimal success.) Probe perturbation effects were also studied, with large variations in velocity discovered when the probe was inserted, due to both local and large-scale effects. This is to be expected in elliptic flows.

Chandran et al. (1984) measured temperature with a coated, 25 micron thermocouple, coincident with LDV data, in a premixed turbulent flame. An average time constant was determined by a cross-spectral analysis technique, using the response from two closely spaced thermocouples to determine the time constants for both. It was claimed that the use of an average time constant does not introduce significant error in the temperature results. Probe interference was minimized by adjusting the probe configuration until the local velocity closely matched the velocity measured when no probe was in the flow.

Elmore et al. (1986) conducted a specific program for dynamic temperature measurement using a dual-junction method with off-line compensation. The dual junctions were used to determine the response characteristics of the smaller thermocouple, which was then corrected. Correction was not done on a point-by-point basis, but rather in the frequency domain.

Essentially, the technique determined the convective heat transfer coefficient as a function of instantaneous frequency, corrected the measured temeprature for convective response in the frequency domain via extensive use of FFT's, then inverse-transformed (when appropriate) to get temporal waveforms. Frequency response was limited to about 250 Hz. The development program was not intended to provide temperatures in reacting regions, so the assumption of air as the working fluid was justified. In this case, the compensation may very well be a function of frequency only. Changes in local chemistry, and catalytic effects, were not addressed. For reacting flows, this technique would probably suffer from the same problems observed by Heitor et al. A finite-difference conduction correction was applied to all data. A Fortran program to provide for compensation and conduction corrections was developed. The design and fabrication details for the probe, and initial testing, were described in an earlier report (Elmore et al., 1983).

In a theoretical investigation, Chandran et al. (1985) used a two time constant model for a premixed flame with an assumed bimodal temperature PDF. A linear response was assumed for each of the two parts of the flow and time constants calculated using assumed flow properties. For an assumed square wave input, the response of the system was modeled and the calculated mean temperature was found. Using only a single time constant generated errors on the order of 10 percent.

Experimental evalutaion of these results in a premixed methane-air flame was conducted with yttria-beryllia coated thermocouples of 25 and 75 micron wire, operated simultaneously. The junction separation was 1 mm. A cross-spectral analysis technique was used to determine the average time constant for each junction, assuming the same environment for both. An RC circuit was tuned to this mean value to accomplish on-line compensation. Comparison to coincident Rayleigh scattering results (with about 1 kHz resolution) showed reasonably good agreement for temporally-resolved temperature, and errors on the order of 20 percent for the

RMS fluctuation and 10 percent for the mean temperature. These errors were predicted by the response model.

Ozem and Gouldin (1989) used an uncoated, 25 micron thermocouple to determine temperature statistics in a turbulent, premixed, methane-air flame. Radiation and conduction were neglected, and response compensation was conducted off-line. The thermocouple output was amplified, low-pass filtered at 20 kHz, and digitized at 40 kHz for transfer to a MicroVax II. An average time constant was used, calculated from a Nusselt number relation rather than an experimentally determined response.

Cambray et al. (1986) attempted to find instantaneous time constants using two chromel-alumel thermocouples of differing diameters in a buoyant turbulent propane diffusion flame. The calculations were performed off-line and were based on the differences in the response gradients of the two thermocouples. In determining these gradients, however, the signal had been "interpolated with straight segments, and the second derivative of the new signal has been interpolated and twice integrated." This manipulation makes it uncertain as to whether the processed signal resembled that of the original. Results were compared with those found using an average time constant and there was found to be little difference.

The bulk of the evidence indicates that time-response compensation is necessary, at least in turbulent flows, if accurate, temporally resolved temperature data are to be obtained. The use of an average time constant does not appear to be sufficient, and using velocity alone as an instantaneous "corrector" to the time constant does not provide much improvement as the response is a function of temperature, velocity, and chemical composition. A two-wire technique is definitely indicated, but frequency domain compensation loses the temporal information required for coincidence with LDV data and still does not account for chemistry. An instantaneous, real-time correction, based on the differential response between two thermocouples of different sizes, is the most promising approach.

THEORY

Performing an energy balance on a spherical thermocouple junction and neglecting radiation and conduction effects gives the standard equation for response:

$$Tc = Tj + \alpha(dTj/dt)$$

where Tc is the instantaneous gas temperature, Tj is the junction temperature, and the time constant is defined as

$$\alpha = (\rho j Cj dj^2)/(6\ k\ Nu)$$

Where ρj is the density of the junction wire, Cj is the specific heat of the junction wire, dj is the junction diameter, k is the gas conductivity, and Nu is the Nusselt number. For turbulent flows, the Nusselt correlation is taken as

$$Nu = C_1 Re^{C_2} Pr^{C_3}$$

where Re is the Reynolds number based on wire diameter and Pr is the gas Prandtl number. Substituting relations for Nu, Re, and Pr into the relation for α:

$$\alpha = [\ (\rho j Cj dj^{2-C_2})/6\ ] \times [\ \mu / (V^{C_2} \rho^{C_2} Pr^{C_3}\ C_1\ k)\ ]$$

Which may be grouped as

$$\alpha = (C_4)(G)$$

where $C_4 = C_4$(wire properties) and G = G(gas properties). Now,

$$Tc = Tj + C_4 G(dTj/dt)$$

Clearly, G is the only unknown in solving for Tc. However, for two thermocouples in the same environment, we can let $Tc_1 = Tc_2$ and solve for G in terms of known or measurable parameters:

$$G = [Tj_2 - Tj_1] \,/\, [\, (C_4\, dTj/dt)_1 - (C_4\, dTj/dt)_2 \,]$$

With a known value for G, the time constants can now be determined and an instantaneous gas temperature obtained. For junction 1 this expression would reduce to:

$$Tc_1 = Tj_1 + [(Tj_2 - Tj_1)(dTj/dt)_1] \,/\, [\, (dTj/dt)_1 - (d_2/d_1)^{2-C_2}(dTj/dt)_2 \,]$$

when both junctions are of the same material, leaving the corrected gas temperature a function only of the diameter ratio between the two thermocouples, the measured temperatures and the temperature gradients.

This is a dynamic calibration and depends upon the responses of the two different junctions in the same environment. The only required information is the value of the coefficient C_2, which has been shown to be about 0.5 but can be experimentally verified.

MODELED DATA

In order to show that it is possible to obtain compensated temperatures based on the technique just outlined, a theoretical response curve was generated to test this technique. By letting $Tc = T_0 \sin\omega t$ and substituting back into the relation $Tc = Tj + \alpha(dTj/dt)$ an expression for the response of each junction Tj could be obtained:

$$Tj(t) = T_0\alpha[(\sin(\omega t)/\alpha) - (\omega\cos(\omega t)) + \omega e^{(-t/\alpha)}] \,/\, [1+(\omega t)^2]$$

By assuming a junction diameter ratio of 0.5 (1:2) and assigning arbitrary values to T_0, ω, and the other physical variables contained within α, the response curve for each junction was calculated. This can be seen in Fig. (1) where the curve of largest amplitude is the actual gas temperature Tc and the next smallest and smallest curves are the response curves of junctions having diameter ratios of 1:2 respectively. These response curves were then fed into a Fortran

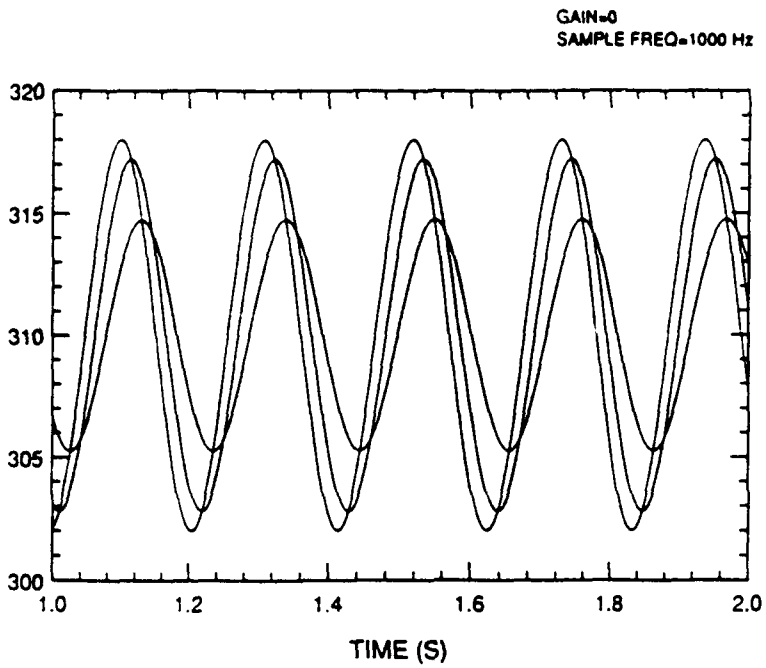


Figure 1 - Theoretical Sinusoidal Gas Temperature With Corresponding Response Signals.

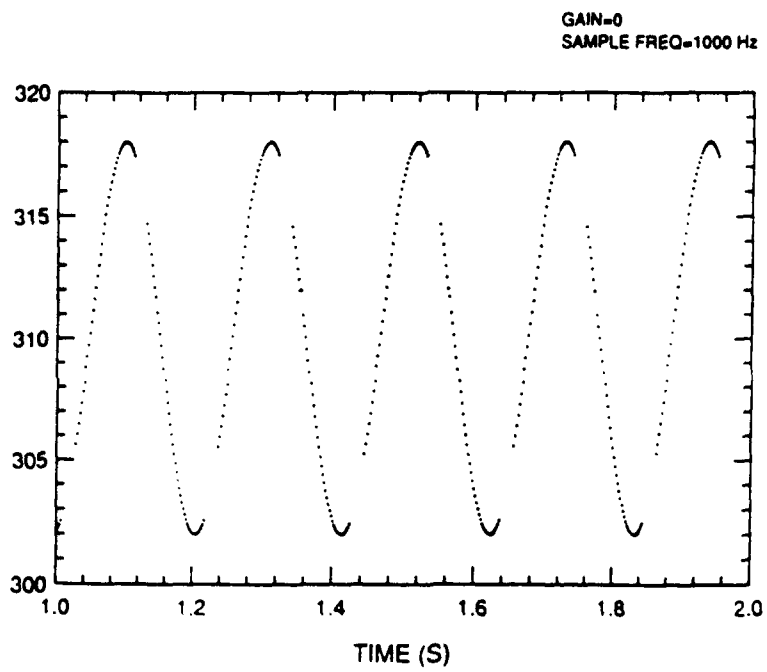


Figure 2 - Compensated Gas Temperatures Based On Sinusoidal Model.

program (Appendix 2) which calculated the gradients at each point and then found the corrected temperatures for both curves. The results, seen in Fig. (2), showed that the corrected temperatures for both response curves were identical and that they matched the actual gas temperature to within 0.1 percent. Moreover, it was found that the time constant, α, could be calculated using only values for the response gradients, the response difference (Tj_2-Tj_1), the diameter ratio, and the value C2. The physical properties of the junctions were not a factor as long as both junctions were of the same material. Similar calculations were performed for various diameter ratios giving similar positive results, thus showing that this technique works well in theory for any given diameter ratio.

A second theoretical model was established to simulate an actual experiment in which two thermocouples of differing diameters at some initial temperature T1 were suddenly exposed to a constant temperature environment T2. The response for each junction to such a step change looks like:

$$Tj(t) = T2 + (T1-T2)\, e^{-(t/\alpha)}$$

For the case where T1=0.2, T2=2.5, and the diameter ratio is 0.5 for both the generation of theoretical responses and the regeneration of the temperature T2, such a compensation was indeed sucessful. However, in order to test the sensitivity of the compensation technique to the diameter ratio, it is possible to try compensating with ratios somewhat different than that of 0.5. When this was attempted, this technique was found to have a great deal of sensitivity to the diameter ratio. Being off by even 1% (eg 0.49) from the actual ratio produced numbers dramatically different from that of the actual gas temperature, T2. In fact, a pattern was seen in the compensation curves as the diameter ratio was changed. When the compensation ratio began below that of the correct ratio (eg 0.25) and was increased, the compensation curve slowly increased and approach the correct gas temperature in an asymptotic fashion as seen in Fig. (3). When the compensation ratio equaled that of the correct ratio, the compensation curve gives the actual gas

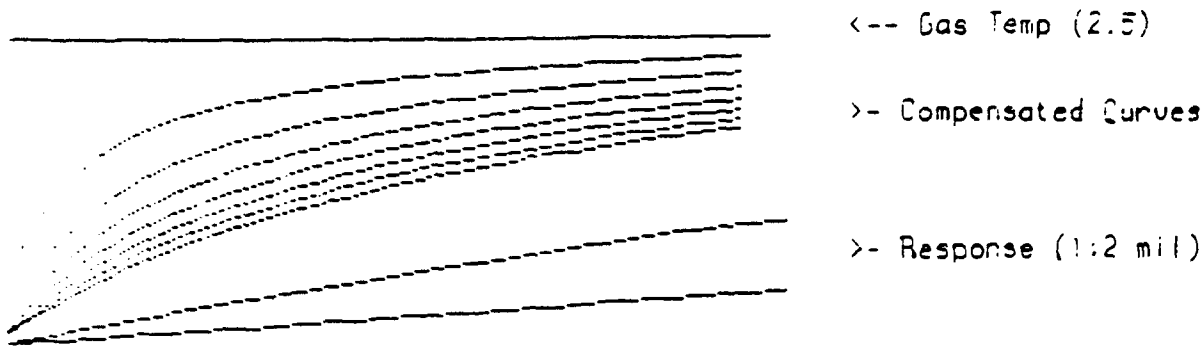


Figure 3 - Results Of Compensation Using Diameter Ratios Less Than Actual Ratio Of 0.5 (Ratios Used: .43 to .49).

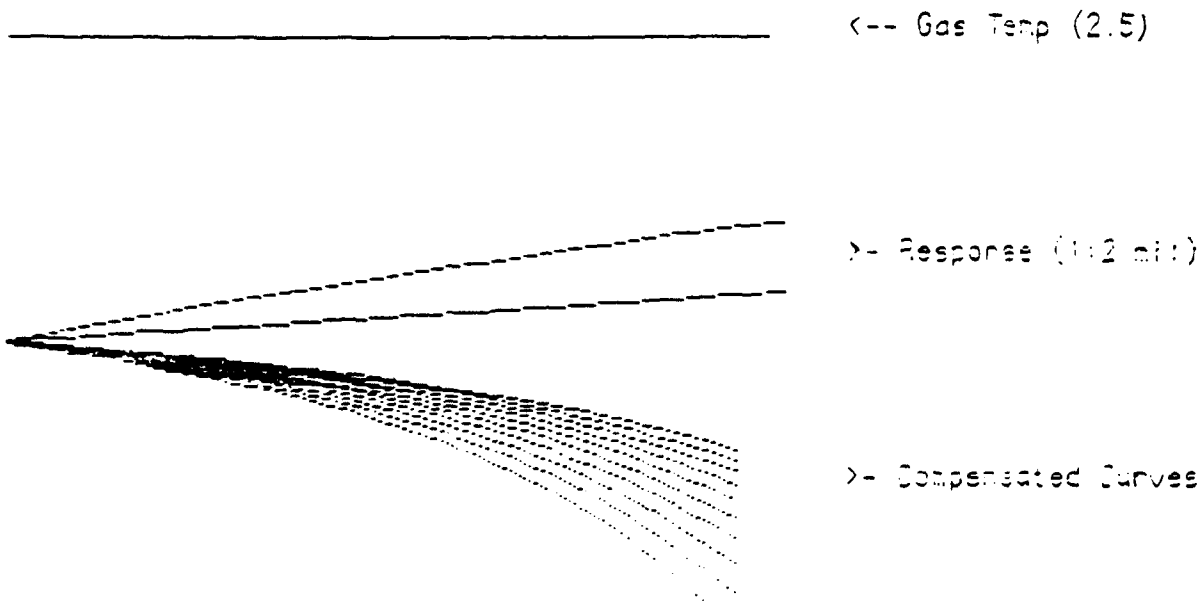


Figure 4 - Results Of Compensation Using Diameter Ratios Above Actual Ratio Of 0.5 (Ratios Used: .70 to .80).

temperature, T2. When the ratio surpased the correct ratio (eg 0.75), the compensation curve jumped from positive to negative values in such a fashion that it began to approach the lower response curve, shown in Fig. (4). All this suggests a means of calibrating a probe to the correct diameter ratio by looking for some sort of convergence to an established gas temperature.

Sensitivity of this method to fluctuations in the response signal was also examined. This was done by adding a small amount of random noise to the theoretical response curves. Incorporating noise on the order of 0.01% of the magnitude of the signal resulted in fluctuations in the response curve on the order of plus or minus 50 percent of the actual temperature, while noise on the order of 0.1% of the signal created very great disturbances which produced non-physical (negative or extremely large) values for the corrected temperature. This suggested that in an experimental situation the signal should have a high signal to noise ratio for correct compensation to occur. It also demonstrated the sensitivity of compensation to the calculated gradients. If the gradients used in compensation deviate even a small amount from their true value, large errors in corrected temperature may result.

EXPERIMENTAL

Data for all experiments was collected using a 386-based computer containing a DT2831-G A/D board with a maximum throughput rate of 250 kHz. The typical data collection rate per channel was 10 kHz. Thermocouple outputs were amplified and filtered using two Preston Model 8300XWB amplifiers with filter cutoffs of 10 Hz, 500 Hz, 1 kHz, 5 kHz, and 10 kHz. A gain of 1000 was typically needed to produce outputs on the order of 1 volt at design temperatures.

The thermocouple wires were made of Pt/Pt-13%Rh and were of1, 2 or 3 mil diameter. Smaller wires would have higher frequency responses but have previously been shown to have unacceptable survivability in the high temperature, high velocity flows that the

two-wire probe is intented to measure. The lead wire was 15 mils in diameter. Probes using 1:2 and 1:3 diameter ratio thermocouples were constructed. The junctions were flame welded using a small hydrogen torch that created a bead which was typically four times the wire diameter and was usually not a perfect sphere. The smallest beads that most resembled spheres were chosen for the construction of each probe. The junction wires were connected to the lead wires using an electrical welder. This proved, through trial and error, to give sufficient strength to withstand flows of approximately 150 ft/s. An attempt was made to place the junctions as close together as possible, but a spatial separation of up to several millimeters often could not be avoided.

Initial attempts at data aquisition sought to move the probe in and out of a methane diffusion flame. This set-up, shown in Fig. (5), used a 200 Watt speaker driver with a 60 oz. magnet to rotate the probe back and forth around a pivot point, thus exposing the junctions to the hot environment at a known frequency. This was moderately successful at exposure frequencies smaller than 20 Hz. However, anything higher would either not provide enough amplitude to move the junction in and out of the flame, or it would simply destroy the probe itself by shaking apart the junctions along with the ceramic insulation. Also, the use of a methane diffusion flame required that catalytic effects be considered, along with possible radiative losses.

The solution to these problems involved placing a dual-junction probe in a stream of pulsed hot air of known temperature in order to provide a series of response curves which could be used to find the instantaneous gas temperature. Figs. (6) and (7) show details of this experiment. Hot air was provided by passing compressed air through an electrical resistance heater which typically supplied values between 300°F and 600°F. The hot flow was then cycled using a rotating steel plate with a hole drilled through it. The test probe and the air supply were then placed on either side of the spinning chopper wheel and a laser and photomultiplier tube (PMT) were used to mark the onset of exposure to the hot air each time the

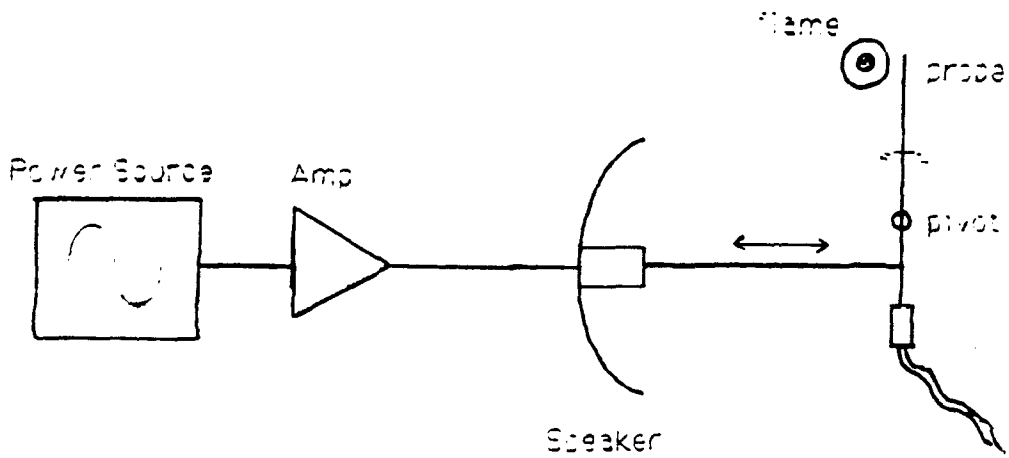


Figure 5 - Experiment To Vibrate Probe Through Flame Using Speaker Driver.

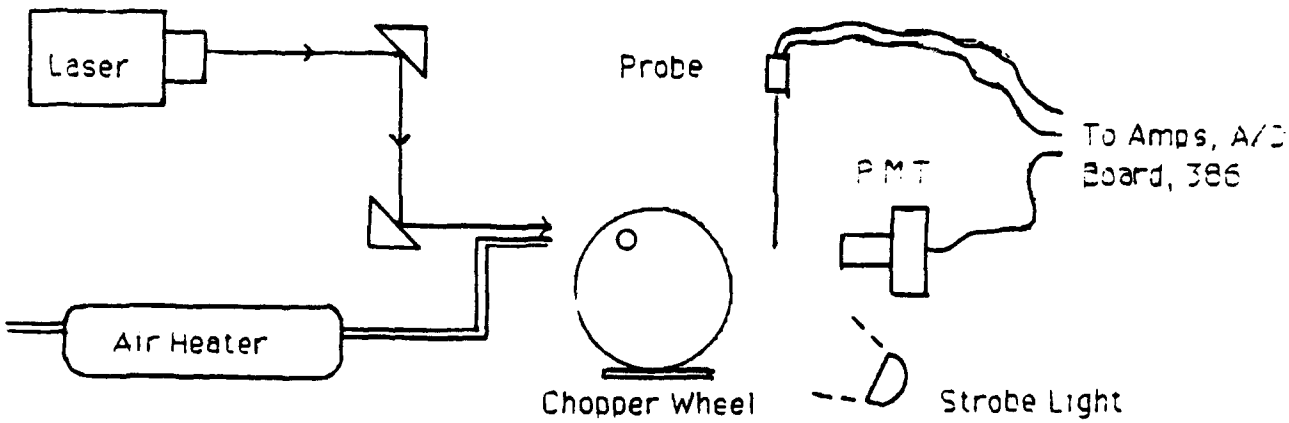


Figure 6 - Experiment To Expose Probe To Hot Air Pulses Using Chopper Wheel And Laser To Mark Onset Of Exposure.

Figure 7 - Photograph Of Chopper Wheel Experiment.

hole passed in front of the probe. Typical wheel speeds, measured with a strobe, ranged between 600 rpm and 3000 rpm.

A minimum of three channels were required for this experiment; one for each thermocouple and one for the PMT signal. In the future, the PMT input channel may be replaced by an LDV input signal. Data was collected at a rate of 10 kHz and was amplified at a gain of 1000 and low pass filtered at 10 kHz.

Fig. (8) shows a typical response curve for a wheel speed of 600 rpm at an air temperature of 460°F. Similarly, Fig. (9) shows another response curve at the same air temperature for a wheel speed of 1200 rpm. It should be noted that wheel speeds of 600 rpm and 1200 rpm correspond to actual driving frequencies (frequencies to which the junctions actually respond) of about 220 Hz and 350 Hz respectively. These numbers are based on an analysis of experimental data to be discussed in the following section.

ANALYSIS

The experimental data used for compensation testing comes from a probe with a measured diameter ratio (assuming spherical junctions) of approximately 0.540 (4.3:8.0 mils) which was held in a pulsed air stream of 300°F (1mV), 460°F (1.8mV), and 600°F (2.6mV) at wheel speeds of 600 rpm and 1200 rpm.

Initially, compensation was attempted on numerous curves at various gas temperatures and wheel speeds using the measured diameter ratio and a linear approximation technique in which the gradients of the response curves were determined by finding the slopes between the points behind and in front of the points being compensated. This simple attempt produced results, shown in Fig. (13), that fluctuated excessively and did not approach any sort of physically realistic profile for the corrected gas temperature. Also, the calculated time constant, α, often was negative using this method. This appears to be due to the inability of the simple linear slope approximation to produce voltage gradients that are accurate enough to overcome the sensitivity of the method to such values. Another

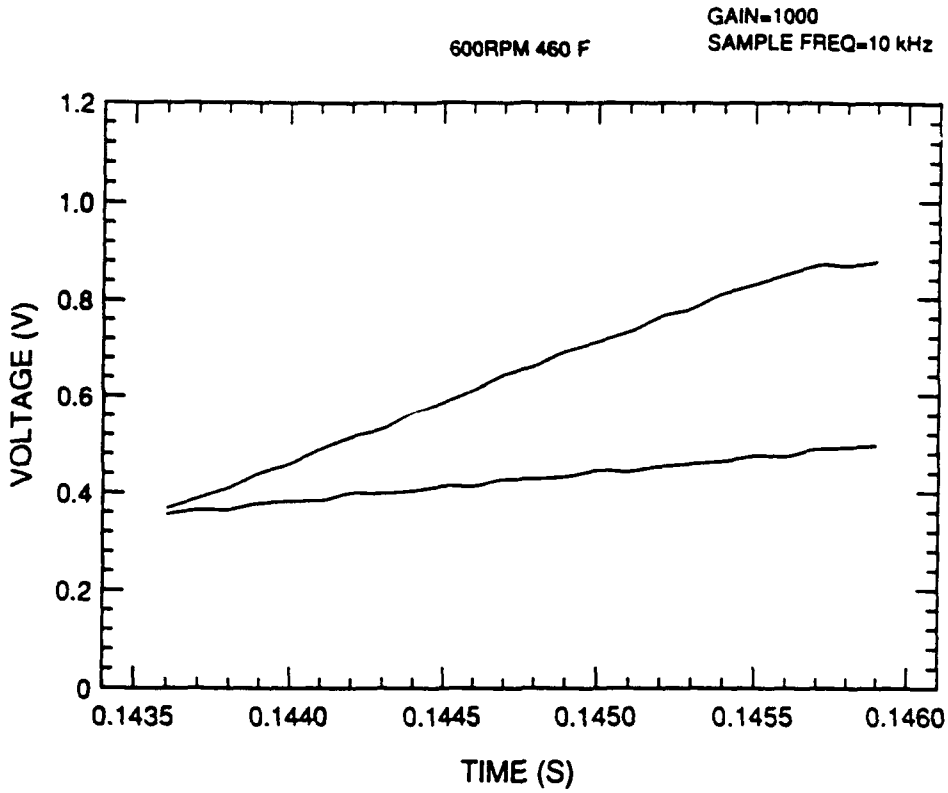


Figure 8 - Typical Response Curves At 600 rpm, 460F.

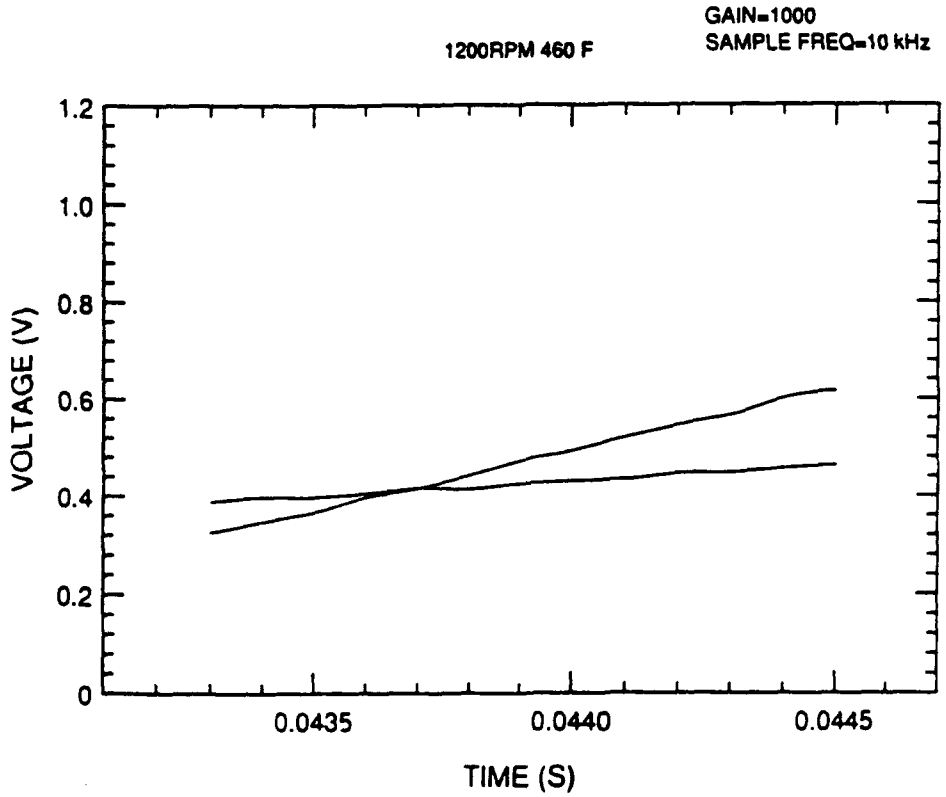


Figure 9 - Typical Response Curves At 1200 rpm, 460F.

problem was that each junction was not a perfect sphere as the theory assumes, so that the measured diameter ratio may have differed from the true "effective" diameter ratio.

An attempted solution to this problem was to take an ensemble average of many curves (50 for a wheel speed of 600 rpm and 100 for 1200 rpm) in order to smooth out the variations in the response signals, and then to try using various diameter ratio values so that a range of ratios giving feasible solutions could be developed for each temperature and wheel speed. If these ranges overlapped, a region of effective diameter ratios would be obtained. This "calibration" of the thermocouple probe is necessary because of an unknown actual gas temperature profile and an unknown effective diameter ratio. The smoothing that occurs in the ensemble averaging is done to try and isolate the 200 Hz to 400 Hz driving frequency to which the probe actually responds. The smaller, high-frequency components are not of interest for these calibration purposes and must be removed to get proper differentiation of the 200 Hz to 400 Hz component when using a linear slope approximation.

A trial-and-error approach of this type was tried on the ensemble averaged data at 460°F (1.8mV) and 600 rpm. The results, as seen in Fig. (10), seem to indicate a possible ratio range of about 0.331 to 0.350 with fluctuations in the compensated temperature on the order of 10% of the measured gas temperature. Likewise, at 460°F and 1200 rpm, an approximate range of about 0.344 to 0.363 was found, as seen in Fig. (11). With these results in mind, compensation of the averaged curves for 600°F (2.6mV) and 300°F (1mV) at 600 rpm were performed, producing a ratio range in both cases of about 0.344 to 0.356. The time constants in all cases also appeared reasonable, showing values on the order of 5 to 10 msec for the smaller junction and 20 to 40 msec for the larger junction. A calculation for the time constant using estimated physical parameters gives a value of about 6 msec for a 4 mil bead and 17 msec for an 8 mil bead.

It was noted in performing these manipulations that there was an overall trend regarding the compensation curves and their behavior when going from a low diameter ratio to a high ratio. This

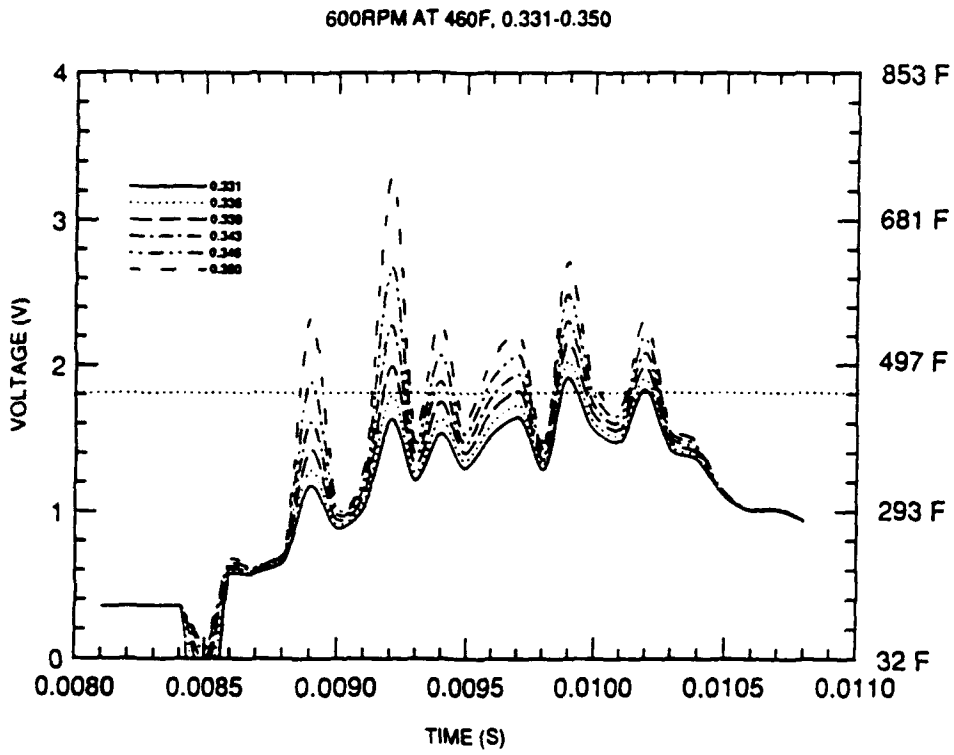


Figure 10 - Compensation On Ensemble Average Curves At 460F, 600 rpm, Diameter Ratios 0.331 to 0.350.

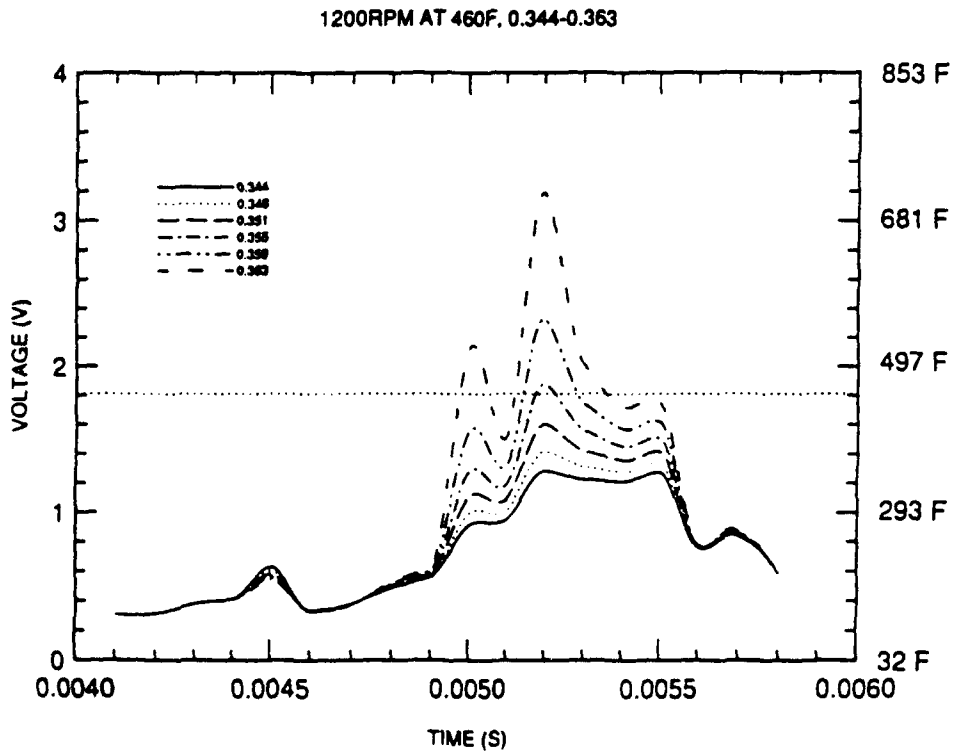


Figure 11 - Compensation On Ensemble Average Curves At 460F,1200 rpm, Diameter Ratios 0.344 to 0.363.

trend was similar to that seen in the theoretical model in that the compensated curve grew larger in value with increasing ratio and eventually "flipped over" and approached the lower response curve at high ratios. However, there was never a point of definite convergence on one value as was seen in the model. Rather, there was a gradual shift in the response curve as the ratio changed, thus necessitating the need for a range of possible solutions.

Using the resulting diameter ratio range, compensation was attempted on single, non-averaged response curves. This, once again, produced fluctuating and unrealistic results. It took an ensemble average of about 20 or more curves to suppress the high-frequency signal fluctuations enough that the derivatives were sufficiently accurate and useful results could be obtained.

Since the compensation method is clearly so dependent on the calculated gradients, the likely approach would seem to be to use a curve fit over the response curve. Second order, third order, and high (10th) order curves fits were tried, but all were initially unsuccessful and gave unrealistic answers. However, it was found that a high order curve fit over the low curvature response portion of the curve (ie the period of time between exposure of the junction to the hot air and when the hot air flow is stopped) produced results that were physically reasonable and that did away with much of the resulting fluctuation.

A high order curve fit was used on 600 rpm averaged curves at 300°F, 460°F, and 600°F to obtain a ratio of approximately 0.331 to 0.338. This is a slightly smaller but narrower range of values than that found using simple linear slopes. Curve fits were also done on single, non-averaged response curves using this range of ratios, and this gave smooth, generally realistic results. Fig. (12) shows a typical set of compensation curves using this range of ratios for a single response at 460°F (1.8mV) at a wheel speed of 600 rpm. This can be compared to Fig. (13) which shows compensation curves for the same response curve without using a high order curve fit to determine the gradients. Fig. (14) shows typical results at a wheel speed of 1200 rpm and 460°F. Compensated temperatures of single curves at 600°F and 300°F at 600 rpm, shown in Figs. (15) and (16)

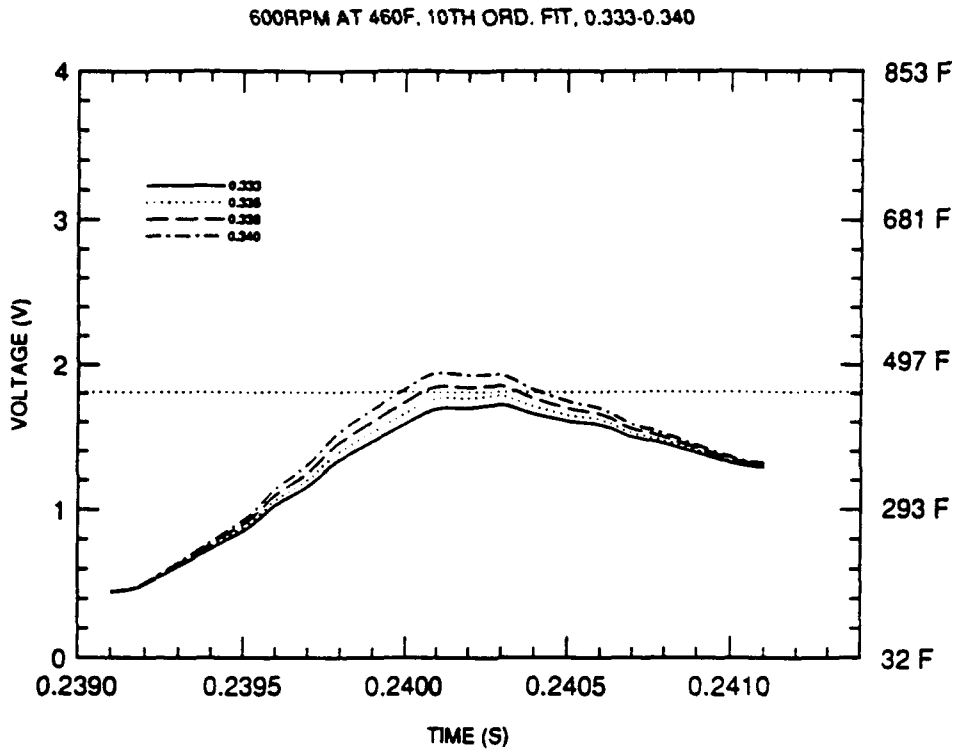


Figure 12 - Compensation On Single, Non-Averaged Curve At 460F, 600 rpm, Diameter Ratios 0.333 to 0.340 Using10th Order Curve Fit.

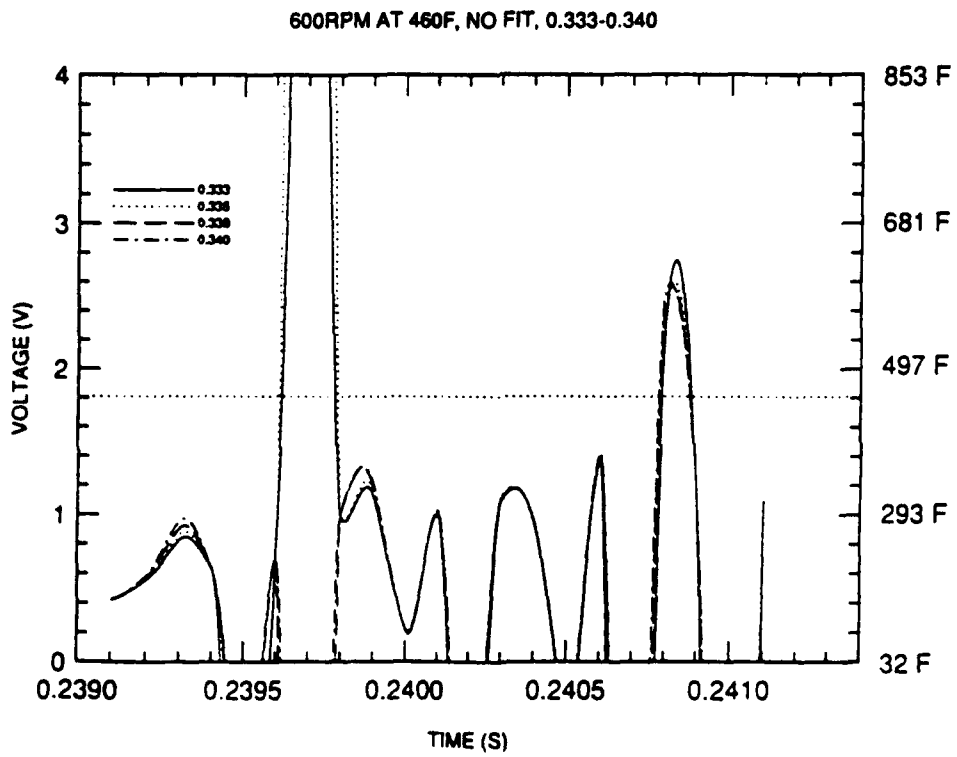


Figure 13 - Compensation On Single, Non-Averaged Curve At 460F, 600 rpm, Diameter Ratios 0.333 to 0.340 Using Linear Slopes.

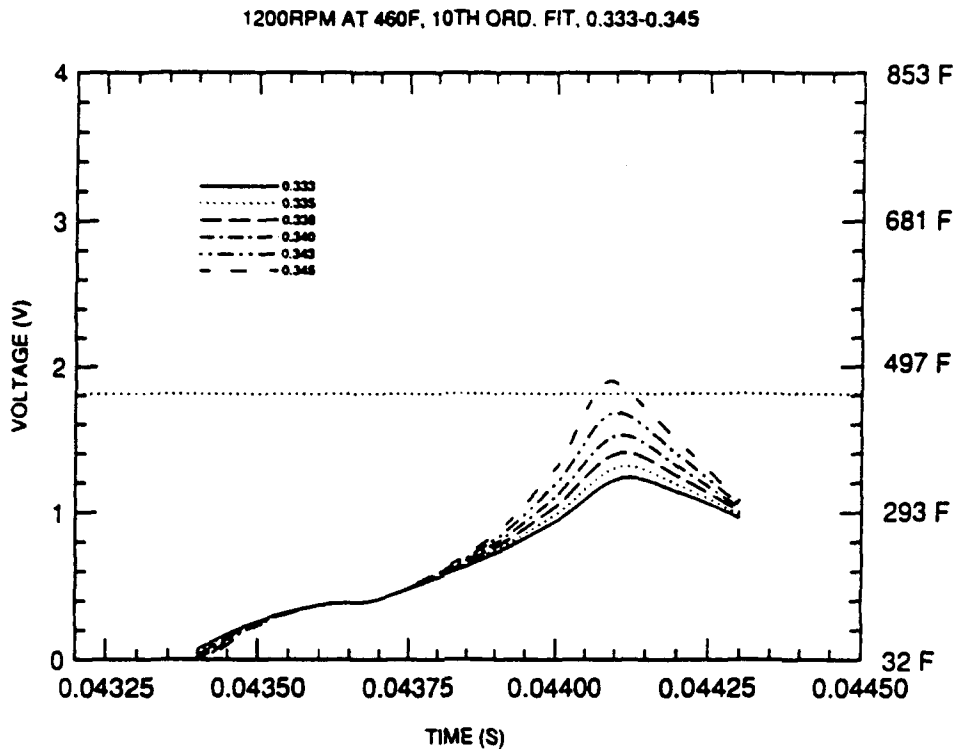


Figure 14 - Compensation On Single, Non-Averaged Curve At 460F, 1200 rpm, Diameter Ratios 0.333 to 0.345 Using10th Order Curve Fit.

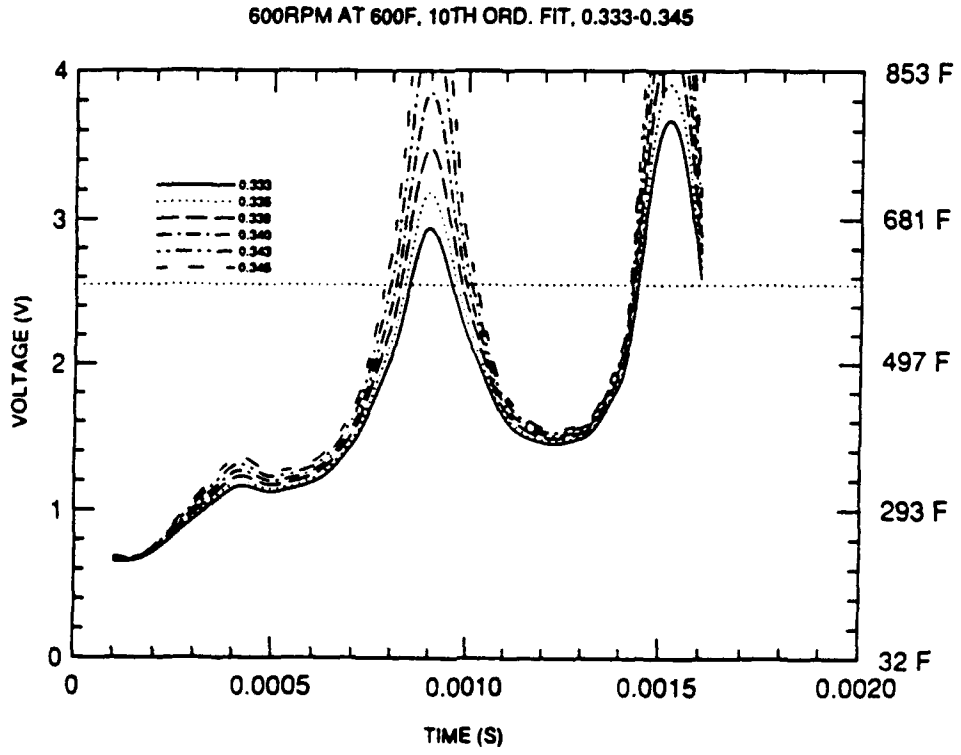


Figure 15 - Compensation On Single, Non-Averaged Curve At 600F, 600 rpm, Diameter Ratios 0.333 to 0.345 Using10th Order Curve Fit.

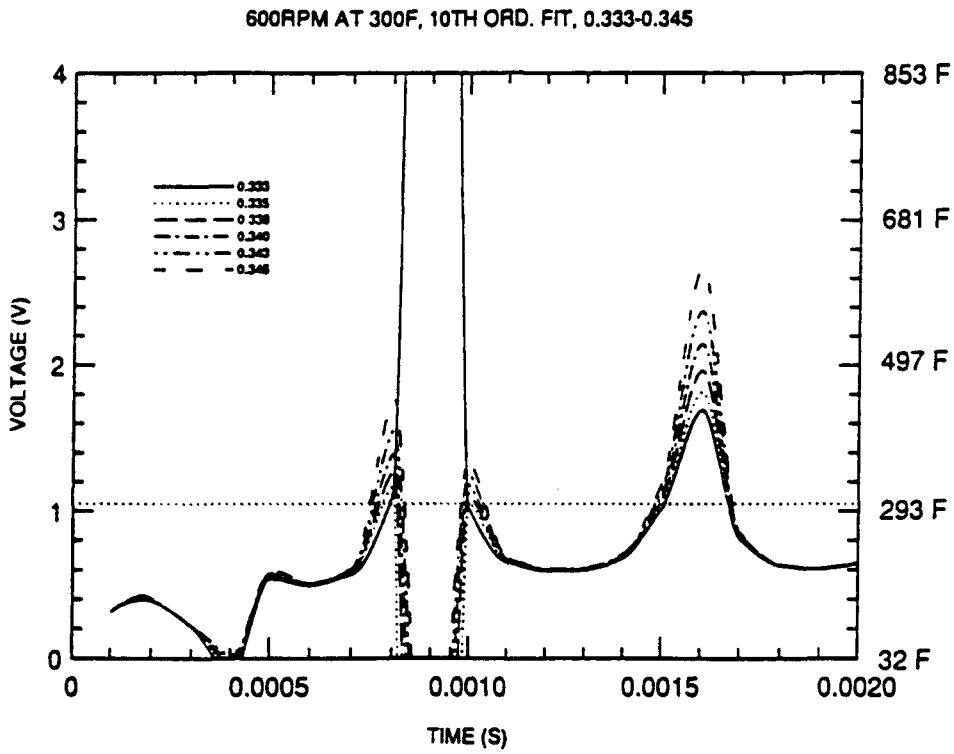


Figure 16 - Compensation On Single, Non-Averaged Curve At 300F, 600 rpm, Diameter Ratios 0.333 to 0.345 Using10th Order Curve Fit.

respectively, demonstrate a different behavior, but are not totally unrealistic. These figures show the compensated gas temperature peaking twice, which may be due to fluid mechanical effects. The large magnitude of these peaks can be attributed to a slight amount of high-frequency "noise" in the signal or to small imperfections in the calculated gradients. All these results correspond to the ratio range of about 0.331 to 0.338 and all give time constants on the order of those predicted.

CONCLUSION

The experimental results indicate that temperature compensation using a two-wire thermocouple probe is possible. By calibrating the probe to a range of effective diameter ratios at one known gas temperature, reasonable values for compensated temperature were achieved at other gas temperatures as well. The range of effective diameter ratios was generally found to be below the measured ratio of 0.538 by about 60%.

The compensated temperature curves showed a strong dependence on the calculated temperature gradients. When these gradients were found through the use of linear approximation, a meaningful range of diameter ratios, about 0.344 to 0.356, could only be obtained using an ensemble average of at least 20 curves which served to smooth out the response curve by eliminating unwanted high-frequency components. The resulting ratios were then applied to single, non-averaged response curves and the compensated temperature curves showed a large degree of fluctuation and did not represent a phsyically realistic solution.

However, when the original curve was filtered to eliminate high-frequency components by fitting a high (10th) order equation, much of this fluctuation disappeared and a slightly smaller but narrower range of diameter ratios, about 0.331 to 0.338, was found. This demonstrated the significance of obtaining an accurate temperature gradient, but at the price of making these slopes more complicated to find. One solution to this may be to use a lower order

curve fit over small portions of the response curves at a time. Another possible idea would be to use a differentiating op-amp circuit to obtain the slopes electronically, although frequency response and stability of such a circuit would have to be considered.

A significant problem remains concerning the unknown nature of the actual gas temperature profile. Since this is not known, an exact solution for the effective diameter ratio cannot be obtained and a range of values must suffice. To avoid this, it may be possible to place a very fine wire temperature sensor in the flow as a calibration device with the two-wire probe. Wires with sufficient frequency response to serve as calibration standards are commercially available. While such wires would not survive the high temperature, high velocity applications intended for the two-wire probe, they could be used as references in less severe conditions such as produced in the chopper-wheel calibration rig. This would provide a means of checking results with a known solution in order to obtain a more precise value for the effective diameter ratio, and thus a more accurate compensation. Eventually, correlation of this method with a CARS system may also be possible.

To conclude, it has been demonstrated that temperature compensation based on the response of two different-sized thermocouples appears possible both in theory and in practice. A method of calibrating thermocouple probes to an effective diameter ratio has been demonstrated and has been shown to give realistic results across several different temperatures and driving frequencies. Future work will concentrate on pinpointing the effective diameter ratio more accurately.

REFERENCES

Abdalla, A. Y., D Bradley, S. B. Chin, and C. Lam, 19th Symposium on Combustion, pp. 495-502, 1982.

Brum, R. D., E. T. Seiler, J. C. LaRue, and G. S. Samuelson, AIAA-83-0334, 1983.

Cambray, P., Vachon, M., Maciaszek, T., and J. C. Bellet, 23rd ASME National Heat Transfer Conference, Denver, CO, August 1986.

Chandran, S. B. S., N. M. Komerath, and W.C. Strahle, 20th Symposium on Combustion, pp. 429-435, 1984.

Chandran, S. B. S., N. M. Komerath, W. M. Grissom, J. I. Jagoda, and W. C. Strahle, Comb. Sci. and Tech., vol. 44, pp. 47-60, 1985.

Cookson, R. A., P. G. Dunham, and J. K. Kilham, Comb. and Flame, vol. 8, pp. 168-170, 1964.

Collis, D. C., and M. J. Williams, J. Fluid Mech., vol. 6, pp. 357-384, 1959.

El Banhawy, Y., S. Sivasegaram, and J. H. Whitelaw, Comb. and Flame, vol. 50, pp. 153-165, 1983.

Elmore, D. L., W. W. Robinson, and W. B. Watkins, NASA CR-168267, 1983.

Elmore, D. L., W. W. Robinson, and W. B. Watkins, NASA CR-179513, 1986.

Fristrom, R. M., and A. A. Westenberg, Flame Structure, pp. 170-174, McGraw-Hill, 1965.

Heitor, M. V., A. M. K. P. Taylor, and J. H. Whitelaw, Exp. in Fluids, vol. 3, pp. 323-339, 1985.

Katsuki, M., Y. Mizutani, and Y. Matsumoto, Comb. and Flame, vol. 67, pp. 27-36, 1987.

Kent, J. H., Comb. and Flame, vol. 14, pp. 279-281, 1970.

Lenz, W., and R. Gunther, Comb. and Flame, vol. 37, pp. 63-70, 1980.

Lockwood, F. C., and H. A. Moneib, Comb. Sci. and Tech., vol. 22, pp. 63-81, 1980.

Lockwood, F. C., and A. O. Odidi, Comb. Inst. European Symp., pp. 507-512, 1973.

Masri, A. R., and R. W. Bilger, 20th Symposium on Combustion, pp. 319-326, 1984.

Masri, A. R., and R. W. Bilger, 21st Symposium on Combustion, pp. 1511-1520, 1986.

Moffat, R. J., Temperature, It's Measurement and Control in Science and Industry, vol. 3, part 2, pp. 553-571, Reinhold, 1962.

Ozem, H. L., and F. C. Gouldin, Paper 36, Eastern Section Of The Combustion Institute, 1989.

Starner, S. H., and R. W. Bilger, 21st Symposium on Combustion, pp. 1569-1577, 1986.

Yamazaki, M., and Ohya, Nenryo Kyokai-Shi, vol. 62, no. 673, pp. 318-326, 1983.

Yoshida, A., and H. Tsuji, 17th Symposium on Combustion, pp. 945-956 1978.

Yule, A. J., D. S. Taylor, and N. A. Chigier, J. Energy, vol. 2, no. 4, pp. 223-231, 1978.

Appendix 1 - The Effects Of Probe Geometry On Surrounding Flow.

The effect of probe geometry on the flowfield surrounding the thermocouple junctions was studied in anticipation of future applications of the probe to a high-speed reacting environment such as that found in a ramjet combustor. This study was carried out in a simple dump, no-swirl combustor using a two-component LDV system. The probe geometries studied, shown in Fig. (17), included two pitot-tube-style configurations in which the U-shaped bend was horizontal (parallel) with the flow and also vertical (perpendicular) to the flow. The third configuration involved a straight probe inserted perpendicular to the flow. Results were obtained at distances of five and ten inches from the combustor outlet and plots of these figures are given in Figs. (18) and (19). There was no significant difference between the three geometries tested, although at five inches the straight probe appeared to have a slight effect by giving values up to 15 percent below those obtained with no probe inserted in the flow.

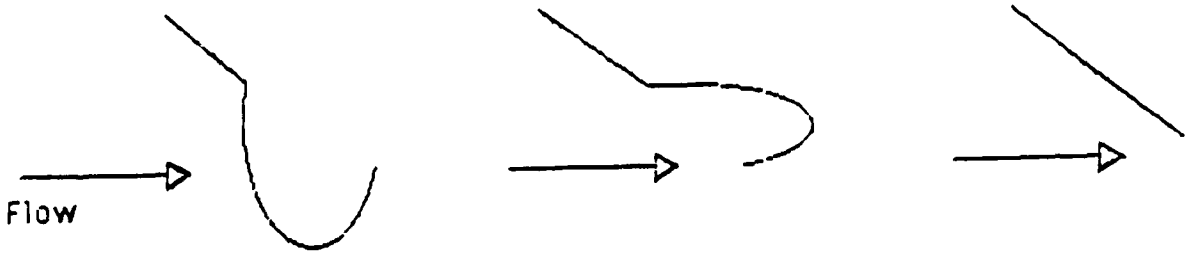


Figure 17 - Various Probe Geometries Tested.

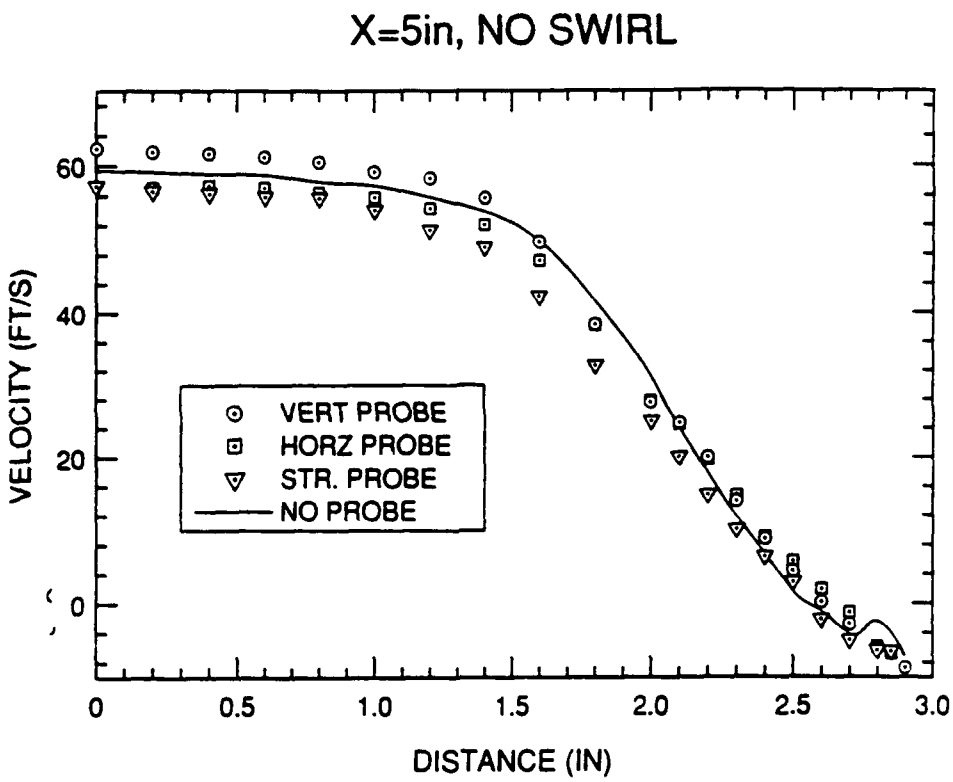


Figure 18 - Probe Flow Effects At 5in In Simple Dump Combustor

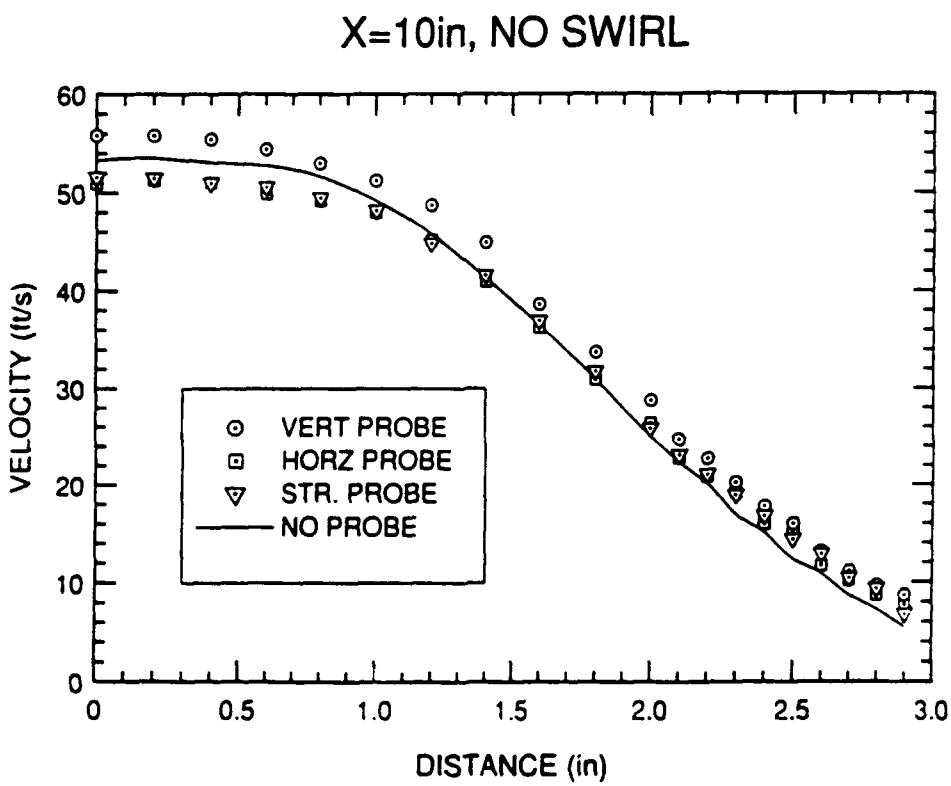


Figure 19 - Probe Flow Effects At 10in In Simple Dump Combustor

Appendix 2 - Fortran Program Used To Perform Compensation Calculations

```
c234567
c      *** Program TCALC,  by Tom Scattergood, 1991
c      *** This program performs compensation calculations on
c      *** response signals using curve fits, linear slopes, and
c      *** various other nifty options.
c
       REAL TIME1A(9000),TIME2A(9000),DVDTA(9000),
      *TIME1B(9000),TIME2B(9000),DVDTB(9000),TSEGA(500),
      *TSEGB(500),TEQL(500),ACOE1(20),ACOE2(20),BCOE1(10),BCOE2(10)
       REAL T1A,T1B,T2A,T2B,RHO1,RHO2,CP1,CP2,C2,DIA1,DIA2,C41,C42
      *G,DVDT1,DVDT2,V1,V2,D,GNEG,GPOS,XDVDT1,XDVDT2,INC,
      *MAX,MIN,AVG1,AVG2,SIGAVG,XX,VGAS,DUMMY
       INTEGER X,CNTRA,CNTRB,Y,Z,Q,W,FLAGA,FLAGB,E,SHOW,NOPTS
       INTEGER PICK1,PICK2,PICK3,PICK4,DISC,MARK,PTSACT,POWER
       CHARACTER*12 CHAN1M,CHAN2M,TMPDAT,SLPDAT,SETUP,DATE,RANGE,
      *YORNO

       PICK1=0
       PICK2=0
       PICK3=0
       PICK4=0
       MARK=0
       DISC=0
       MAX=0
       MIN=0
       XX=0
       AVG1=0
       AVG2=0
       SIGAVG=0
       PTSACT=0
       OPEN(3,FILE='A10HZC3.COE')
       DO 788 POWER=1,11,1
        READ(3,*)ACOE1(POWER)
788    CONTINUE
       DO 789 POWER=1,11,1
        READ(3,*)ACOE2(POWER)
789    CONTINUE
       CLOSE(3)
       DATE='11/20/91'
       CHAN1M='10AVG1ML.GRA'
       CHAN2M='10AVG2ML.GRA'
       TMPDAT='10HZD.TMP'
       SLPDAT='A10HZC3.SLP'
       RANGE=' '
       YORNO=' '
       NOPTS=22
       INC=.0001
       DIA1=2.72
       DIA2=8.0
       WRITE(*,'(A14,A14,A14,A14,I6,2X,F5.4)')CHAN1M,CHAN2M,TMPDAT,
      *SLPDAT,NOPTS,INC
       WRITE(*,'(A21)')'Change current setup?'
       READ(*,'(A1)')SETUP
       IF (SETUP.EQ.'Y'.OR.SETUP.EQ.'y') GOTO 1
       GOTO 2
1      WRITE(*,'(A1)')' '
       WRITE(*,'(A21)')'ENTER 1 MIL CHANNEL: '
       READ(*,'(A12)')CHAN1M
       WRITE(*,'(A21)')'ENTER 2 MIL CHANNEL: '
       READ(*,'(A12)')CHAN2M
```

```
      WRITE(*,'(A21)')'ENTER DATA FILENAME: '
      READ(*,'(A12)')TMPDAT
      WRITE(*,'(A28)')'ENTER POINT-SLOPE FILENAME: '
      READ(*,'(A12)')SLPDAT
      WRITE(*,'(A18)')'ENTER # DATA PTS: '
      READ(*,'(I6)')NOPTS
      WRITE(*,'(A31)')'ENTER SMALLEST TIME INCREMENT: '
      READ(*,'(F5.4)')INC
2     WRITE(*,'(A48)')'==============================================='
      WRITE(*,'(A37)')'                    SELECT ONE          '
      WRITE(*,'(A1)')' '
      WRITE(*,'(A28)')'(1) Tact AT EVERY POINT      '
      WRITE(*,'(A46)')'(2) Tact ALONG MATCHING TRANSITIONAL INTERVALS'
      READ(*,'(I1)')PICK1
      WRITE(*,'(A1)')' '
      WRITE(*,'(A28)')'(1) LEAVE SIGN OF G AS IS    '
      WRITE(*,'(A28)')'(2) TAKE ABSOLUTE VALUE OF G'
      READ(*,'(I1)')PICK2
      WRITE(*,'(A1)')' '
      IF (PICK1.EQ.1) GOTO 3
      WRITE(*,'(A  )')'(1) USE AVERAGE SLOPE'
      WRITE(*,'(A  )')'(2) USE POINT SLOPE'
      READ(*,'(I1)')PICK3
      WRITE(*,'(A1)')' '
3     WRITE(*,'(A37)')'(1) DISCARD POINTS OF OPPOSITE SLOPES'
      WRITE(*,'(A22)')'(2) IGNORE SLOPE SIGNS'
      READ(*,'(I1)')PICK4
      WRITE(*,'(A1)')' '
      WRITE(*,'(A14)')'SET RANGE(Y/N)?'
      READ(*,'(A1)')YORNO
      IF (YORNO.EQ.'N'.OR.YORNO.EQ.'n') GOTO 4
      WRITE(*,'(A24)')'GIVE MAX VALUE(XXX.XXX):'
      READ(*,'(F7.3)')MAX
      WRITE(*,'(A24)')'GIVE MIN VALUE(YYY.YYY):'
      READ(*,'(F7.3)')MIN
      WRITE(*,'(A33)')'DISCARD PTS OUTSIDE THESE VALUES?'
      READ(*,'(A1)')RANGE
4     WRITE(*,'(A40)')'========================================='
      WRITE(*,'(A13)')'Working......'

      OPEN(1,FILE='TRACE.DAT')
      OPEN(4,FILE=CHAN1M)
      OPEN(7,FILE=CHAN2M)
      OPEN(8,FILE=TMPDAT)
      OPEN(2,FILE=SLPDAT)
      WRITE(8,'(A10,A12)')'FILENAME: ',TMPDAT
      WRITE(8,'(A6,A8)')'DATE: ',DATE
      WRITE(8,'(A14,A14,A14,A14)')'FILES USED: ',CHAN1M,
     *CHAN2M,SLPDAT
      IF (PICK1.EQ.1) WRITE(8,'(A18)')'  Every point used'
      IF (PICK1.EQ.2) WRITE(8,'(A18)')'  Intervals only  '
      IF (PICK2.EQ.1) WRITE(8,'(A18)')'  ABS(G) not used '
      IF (PICK2.EQ.2) WRITE(8,'(A18)')'  ABS(G) used     '
      IF (PICK3.EQ.1) WRITE(8,'(A18)')'  Average slopes  '
      IF (PICK3.NE.1) WRITE(8,'(A18)')'  Point slopes    '
      IF (PICK4.EQ.1) WRITE(8,'(A23)')'  Discarded opp. slopes'
      IF (PICK4.EQ.2) WRITE(8,'(A23)')'  Ignored opp. slope   '
      IF (YORNO.EQ.'Y') WRITE(8,'(A13,F7.3,A13,F7.3)')'RangeMAX: ',
     *MAX,'RangeMIN: ',MIN
      IF (RANGE.EQ.'Y') WRITE(8,'(A17)')'Points discarded'
```

```
      WRITE(8,'(A6,F4.2,2X,A6,F4.2)')'DIA1: ',DIA1,'DIA2: ',DIA2
      WRITE(8,'(A25)')'=========================='
      WRITE(8,7)'Time','a1','a2','Vcorr1','Vcorr2'
7     FORMAT(2X,A4,9X,A2,9X,A2,6X,A6,5X,A6)

      REWIND(4)
      REWIND(7)
      REWIND(2)

      RHO1=19.6
      RHO2=RHO1
      CP1=.6
      CP2=CP1
      C2=.5
      C41=(RHO1*CP1*DIA1**(2-C2))/4
      C42=(RHO2*CP2*DIA2**(2-C2))/4
      GNEG=0
      GPOS=0
      CNTRA=0
      CNTRB=0
      X=1

      IF (PICK1.EQ.1) W=NOPTS-1
      IF (PICK1.EQ.1) CNTRA=2
      IF (PICK1.EQ.1) GOTO 50

10    READ (4,*) TIME1A(X),TIME2A(X),DVDTA(X)
      IF (TIME1A(X).EQ.0.AND.TIME2A(X).EQ.0) GOTO 30
      X=X+1
      CNTRA=CNTRA+1
      SHOW=CNTRA
      GOTO 10

30    X=1
40    READ (7,*) TIME1B(X),TIME2B(X),DVDTB(X)
      IF (TIME1B(X).EQ.0.AND.TIME2B(X).EQ.0) GOTO 50
      X=X+1
      CNTRB=CNTRB+1
      GOTO 40

50    DO 100 X=1,CNTRA,1
        IF (PICK1.EQ.1) GOTO 238
        WRITE (*,59) SHOW
59      FORMAT (I4)
        SHOW=SHOW-1
        Y=1
60      T1A=TIME1A(X)
        T1B=TIME1B(Y)
        IF (T1B.LT.(T1A-.001).OR.T1B.GT.(T1A+.001)) GOTO 70
        T2A=TIME2A(X)
        T2B=TIME2B(Y)
        IF (T2B.LT.(T2A-.005).OR.T2B.GT.(T2A+.005)) GOTO 70
        IF (T1A.GT.T2A) GOTO 70
        IF (T1B.GT.T2B) GOTO 70
        XDVDT1=DVDTA(X)
        XDVDT2=DVDTB(Y)
        IF (PICK4.EQ.2) GOTO 62
        IF (XDVDT1.LT.0.AND.XDVDT2.GT.0) GOTO 70
        IF (XDVDT1.GT.0.AND.XDVDT2.LT.0) GOTO 70
```

```
62        Z=1
          DO 333 D=T1A,T2A,INC
            TSEGA(Z)=D
            FLAGA=Z
            Z=Z+1
333       CONTINUE

          Z=1
          DO 334 D=T1B,T2B,INC
            TSEGB(Z)=D
            FLAGB=Z
            Z=Z+1
334       CONTINUE
          W=0

          DO 337 Z=1,FLAGA,1
            DO 335 Q=1,FLAGB,1
              DIFF=ABS(TSEGA(Z)-TSEGB(Q))
              IF (DIFF.LT..0001) GOTO 336
335         CONTINUE
          GOTO 337
336       W=W+1
          TEQL(W)=TSEGA(Z)
337       CONTINUE


238       DO 340 E=1,W,1
            WRITE(*,'(I4)')E
338         READ (2,*) XTIME,V1,V2,DVDT1,DVDT2
            XT=XTIME
            IF (PICK3.EQ.1) DVDT1=XDVDT1
            IF (PICK3.EQ.1) DVDT2=XDVDT2
            IF (PICK1.EQ.1) GOTO 339
            DIFF=ABS(TEQL(E)-XTIME)
            IF (DIFF.LT..0001) GOTO 339
            GOTO 338
339         IF (DVDT1.EQ.0.AND.DVDT2.EQ.0) GOTO 340
             DVDT1=0
             DVDT2=0
             DO 434 POWER=1,11,1
             DVDT1=ACOE1(2)+(2*ACOE1(3)*XT)+(3*ACOE1(4)*XT**2)+
     *(4*ACOE1(5)*XT**3)+(5*ACOE1(6)*XT**4)+(6*ACOE1(7)*XT**5)+
     *(7*ACOE1(8)*XT**6)+(8*ACOE1(9)*XT**7)+(9*ACOE1(10)*XT**8)+
     *(10*ACOE1(11)*XT**9)
            DVDT2=ACOE2(2)+(2*ACOE2(3)*XT)+(3*ACOE2(4)*XT**2)+
     *(4*ACOE2(5)*XT**3)+(5*ACOE2(6)*XT**4)+(6*ACOE2(7)*XT**5)+
     *(7*ACOE2(8)*XT**6)+(8*ACOE2(9)*XT**7)+(9*ACOE2(10)*XT**8)+
     *(10*ACOE2(11)*XT**9)
434          CONTINUE
            WRITE(1,'(F6.4,2X,F9.5,3X,F9.5)')XT,DVDT1,DVDT2
            G=(V2-V1)/((C41*DVDT1)-(C42*DVDT2))
            IF (PICK2.EQ.2) G=ABS(G)
            CONST1=C41*G*DVDT1
            CONST2=C42*G*DVDT2
            TCORR1=CONST1+V1
            TCORR2=CONST2+V2
            IF (DVDT1.LT.0.AND.DVDT2.GT.0.AND.PICK4.EQ.1) GOTO 340
            IF (DVDT1.GT.0.AND.DVDT2.LT.0.AND.PICK4.EQ.1) GOTO 340
            MARK=0
            IF(TCORR1.GT.MAX.OR.TCORR1.LT.MIN) MARK=1
```

```
          IF(TCORR2.GT.MAX.OR.TCORR2.LT.MIN) MARK=1
          IF(MARK.EQ.1.AND.YORNO.EQ.'Y') DISC=DISC+1
          IF(MARK.EQ.1.AND.RANGE.EQ.'Y') GOTO 340
          IF (G.LT.0) GNEG=GNEG+1
          IF (G.GT.0) GPOS=GPOS+1
          WRITE (8,777) XTIME,(G*C41),(G*C42),TCORR1,TCORR2
777       FORMAT (F7.4,3X,F9.5,2X,F9.5,2X,F9.5,2X,F9.5)
          PTSACT=PTSACT+1
          AVG1=AVG1+TCORR1
          AVG2=AVG2+TCORR2
340     CONTINUE
        IF(PICK1.EQ.1) GOTO 887
        GOTO 100
70      Y=Y+1
        IF (Y.GT.CNTRB) GOTO 100
        GOTO 60
100   CONTINUE
887   WRITE (*,888)'-',GNEG,'+',GPOS
888   FORMAT (A1,F6.1,2X,A1,F6.1)
      WRITE (*,889)'There were ',DISC,' points out of range.'
      WRITE (8,889)'There were ',DISC,' points out of range.'
889   FORMAT(A11,I4,A21)
      AVG1=AVG1/PTSACT
      AVG2=AVG2/PTSACT
      SIGAVG=SIGAVG/PTSACT
      WRITE(*,890)'Signal 1 average was ',AVG1
      WRITE(8,890)'Signal 1 average was ',AVG1
      WRITE(*,890)'Signal 2 average was ',AVG2
      WRITE(8,890)'Signal 2 average was ',AVG2
      WRITE(*,890)'     Gas average was ',SIGAVG
      WRITE(8,890)'     Gas average was ',SIGAVG
890   FORMAT(A21,F7.3)
      END
```

Report # 50
210-11MG-098
Prof. Kaveh Tagavi
No Report Submitted

DROPLET DISTRIBUTIONS FROM THE BREAKUP OF A CYLINDRICAL LIQUID JET

L. P. Chin*, T. Jackson^, P. G. LaRose*, J. Stutrud^,
G. Switzer**, R. S. Tankin*

Abstract

A phase/Doppler particle analyzer is used to measure the size and velocity distributions of the droplets generated by the disintegration of a cylindrical liquid jet. This type of liquid jet breakup is commonly called Rayleigh breakup. Metered liquid flow rates agree with the rates computed from the droplet merasurements made with the phase/Doppler particle analyzer. The maximum entropy principle is used to predict the droplet size and velocity distributions. The constraints imposed in this model involve conservation of mass, momentum, surface energy, and kinetic energy. Agreement between measurements and predictions is very good.

I.Introduction

The process of producing droplets by the breakup of a cylindrical liquid jet is commonly called **Rayleigh Breakup** (1879,1882). The use of a stimulated mechanism (forcing function) has been widely employed by many researchers to obtain uniform droplets. Such droplets are essential in calibrating spray sizing instruments and in studying the fundamental aspects of droplet combustion. For a stream of droplets with uniform size and velocity, the only mathematical representation of droplet distribution is a coupled Delta function $\delta(D-D_o)\delta(U-U_o)$. D_o denotes the uniform droplet size and U_o denotes the uniform droplet velocity. This distribution is not generally measured - even when experiments are carefully conducted in a vibration isolated environment. Unforced disturbances and complicated nonlinear breakup mechanism normally produce a stream of droplets deviating slightly from the ideal

* Northwestern University; ^ United States Air Force; ** Systems Research Laboratories, Inc.

distribution. At times such a jet, as a result of nonlinear instability, also has a propensity to produce relatively small "satellite" droplets interspersed among the main droplets (Bogy, 1979). The existence and behavior of these satellite droplets depends on the initial flow conditions and the physical properties of the liquid jet. The stream of droplets with satellites must be described by a bimodal distribution function. In this study, we will not consider flows where satellite droplets exist. To control the production of satellite droplets which is important for the jet printer operation, several studies have been made to understand the formation mechanism at various stimulating disturbances (Goedde & Yuen, 1970; Pimbley & Lee, 1977; Chaudhary & Maxworthy, 1980 and Bousfield et al., 1990, etc.). For example, Pimbley and Lee (1977) obtained a map showing the behavior of satellite droplets under different stimulating conditions. There is a region in this map where the satellite droplet can be totally eliminated. In this study, the measurements recorded by phase/Doppler particle analyzer (PDPA) of an unforced cylindrical jet (as well as photographs) show that there are no satellite droplets produced over the range of operation under consideration.

With a PDPA, it is possible to measure both the droplet velocity and size. The PDPA is a single point scattering technique, making a measurement of each droplet as it passes through a small probe volume. In these measurements the attempts and validations are recorded. It is important that the acceptance ratio (validations/attempts) be close to one; otherwise, the cause of the rejections should be determined. In many complicated sprays (hollow cone, swirl atomizers), an acceptance rate of 60% or less, over portions of the spray, is often obtained. The advantages of using a small diameter cylindrical jet, from an experimentalist's point of view, are twofold: (1) the acceptance rate in the PDPA measurements is very high - approximately 100%. (2) the probe area is not a variable in the problem.

The concept of information entropy was developed by Claude Shannon (1948), and Jaynes (1957) later extended this concept into the now well-known

method of maximum entropy formalism. This formalism can be applied to problems which involve probability, i.e., where insufficient information is available to obtain exact solutions. Tribus (1961) used the principle of this formalism in thermodynamics and showed that the concepts of heat and temperature in thermodynamics could be defined through the formalism of maximum entropy. Haken's book on synergetics (1978) established links between dynamical systems theory and statistical theory with information theory as its basis. This maximum entropy formalism allows one to determine the probability distribution functions for complex systems in physics, chemistry, biology, as well as in many other disciplines by measuring relatively few average (macroscopic) quantities. In the present study, we will confine our attention to its application to liquid sprays in order to predict the droplet size and velocity distributions in sprays. Since the application to these kinds of problems has been adequately discussed by several researchers - Kelly (1976), Sellens & Brzustowski (1985, 1986) and Li & Tankin (1987, 1988, 1989), it will not be necessary to develop the background materials once again.

The primary purpose of this study is to compare the theoretical predictions based on maximum entropy principle concerning droplet size and velocity distributions with the experimental results. In the process of making these comparisons, estimates will be made of the source terms that appear in the conservation equations. From a computational point of view, there is an advantage in using the cylindrical liquid jet instead of a hollow or solid cone spray - reasonably accurate estimates can be made of the source terms associate with these constraints. The pattern of droplets formed after breakup in the Rayleigh problem is different from the regular hollow cone or solid cone spray system. Therefore, the problem is formulated differently than in previousanalyses.

II.Experiment

A critical component in this experiment is the hole that forms the orifice plate. In these experiments the hole is formed by an electroforming process that is commonly used to make bimetal masks for photolithography. With this process precise openings of various shapes (in this case, circular) and sizes (in this case, 50 microns) can be made in a thin metal plate such as nickel. The nickel metal is bonded to a layer of another metal for strength. This technique is used to construct contact exposure masks in the production of printed circuit boards (see Dressler & Kraemer, 1990).

The liquid used in this study is distilled water with the following additives to prevent corrosion at the orifice: ethylene glycol 1%, sodium nitrite 0.1% , borax 0.1% and traces of bacterial inhibitors. The liquid is forced through the orifice by applying a constant air pressure to a liquid reservoir (see Figure 1). By changing the air pressure, the flow rate is changed. The water exiting the orifice forms a cylindrical jet that is approximately 50 microns in diameter. This cylindrical jet breaks up into droplets. It is desirable to measure the droplets as close as possible to the breakup region, thus reducing aerodynamic effects (drag); but beyond the point where the droplets are oscillating (non spherical). Figure 2 is a typical photograph that was taken through a microscope using a strobe as a light source (100 microsecond flash). In this photograph, the droplets as well as a fringe pattern is seen. The fringe pattern, where measurements were made, is due to the laser beams used to determine simultaneously the droplet velocity and size. It can be seen from this photograph that the droplets are spherical at the point where the measurements were taken.

Droplets are sized with an Aerometrics Inc., two-color, four-beam PDPA; although only one color (green - 5145°A) was used in these experiments. This instrument has a probe volume of 0.002 mm^3. It sizes a droplet by measuring its radius of curvature. The instrument is configured like a standard laser Doppler anemometer; the beam pair yields a measurement of the axial component of the droplet velocity. 5,000 droplets

are measured in each test run. Collection times for this many samples are of the order of four seconds. Sample sizes as large as 10,000 droplets were taken before determining that 5,000 samples are more than sufficient to yield stable mean and RMS statistics.

One could have used microscopic photography as shown in Figure 2 and with double exposures and multiple pictures obtained the size and velocity distributions. There are two reasons for using the PDPA unit. First, in an earlier study we examined a complex spray from a pressure atomizer and found the metered flow rate was an order of magnitude greater than the flow rate obtained from the PDPA measurements. Thus we decided to use a simple spray system - a jet that breaks up into a single column of drops - to test the accuracy of our system. The PDPA measurements agreed with the metered flow rate as well as with the sizes estimated from photographs. This will be discussed later. Second, when the distributions are relatively broad (as they were in some cases), analysis of many such photographs (in the hundreds or thousands) would be required to obtain accurate velocity and size distributions. This would have been tedious and time consuming.

Several constraints inherent in the PDPA must be addressed. First, the measured droplets must be spherical. This requires that the measurement station be sufficiently far from the break up region so that droplets are not oscillating (non spherical). Second, the PDPA has a dead time of 16 microseconds associated with each measured droplet. Another droplet entering the probe volume during this dead time will not be measured and may prevent the measurement of the first droplet. In these experiments, the stream of droplets is not dense; thus, there is little probability of having a droplet entering the probe volume during this dead time. These two considerations influence the percentage of valid signals versus the total attempts. That is, the PDPA attempts to process all Doppler signals. It performs checks on the quality of each signal and rejects those which exceed certain limits. In these experiments, there were seven different flow rates analyzed; the valid signals are almost 100% of the signals collected. For example, in one test there are 5022 attempts and 5022 validations - yielding an

acceptance ratio of 100%; in the experiment having the worst acceptance ratio there were 5492 attempts and 5004 validations - yielding an acceptance ratio of 91%

The experimental data for size distribution is normalized by the mass mean diameter, D_{30}. This value is determined from PDPA measurements. The droplet velocity is normalized by the axial velocity at the orifice, U_{jet}. This value is determined by dividing the measured liquid flow rate by the cross sectional area of the cylindrical jet issuing from the nozzle (50 microns diameter). The diameter of the cylindrical jet was verified by magnifying the jet with a microscope and projecting an image onto a screen - from which measurements could be made. Also, direct examination of the cylindrical jet with a microscope with a built in measuring reticle yielded the same result. The cylindrical jet diameter is 50 microns.

III. Theory

As stated, detailed formulation and derivation of the maximum entropy principle have been published (Li and Tankin, 1989), and will not be reiterated here. However there are two significant modifications: (1) Instead of combining the kinetic energy and the surface energy into one energy constraint, these two energy constraints were separated. The need for this will be described later. (2) The control volume chosen here is the region of liquid jet between the orifice plate (nozzle) and the location downstream where the droplets were measured by the PDPA. That means the distribution obtained from this formulation will be the droplet distribution at the same location measured, not the breakup region of liquid jet normally used by Li and Tankin (1989). The reason we chose the measuring position as the downstream boundary is that, later, we will make use of some measured mean values to estimate the source terms. In addition, the aerodynamic effects on the droplet velocities are negligible.

A schematic drawing for the breakup of a liquid jet and it's control volume is shown in Figure 3. The control volume begins at the outlet of orifice (nozzle) and extends far enough downstream so that all of the droplet

oscillations have damped out. The downstream boundary is the location where the measurements were made. The liquid jet which enters the control volume through surface 1 has a diameter D_{jet}, velocity U_{jet}, density ρ, and surface tension σ. The droplets which emerge from the control volume through surface 2 have diameter D, velocity U, and arrive at a frequency $\dot{n}$ (number/time). The surface tension and density are assumed to be unchanged due to the breakup.

By applying maximum entropy formulation, the most probable droplet distribution due to the breakup of cylindrical liquid jet is obtained based on the well known conservation laws involving a few physical quantities. In this study, these physical quantities include mass flux, axial momentum flux, kinetic energy flux and surface energy flux which are conserved in the control volume. Due to the existence of individual droplets crossing surface 2; the mass, momentum and energy leaving the control volume are not constant but occur in a temporal oscillating manner whose frequency depends on $\dot{n}$. Thus, there are changes of these quantities (mass, momentum, kinetic energy, and surface energy) within the control volume with time. To maintain a steady state formulation, the conservation equations employed here will be derived using a time averaged base.

IIIa.Conservation of mass flux, momentum flux and energy fluxes :

If the droplets are produced in a saturated air environment, there is no mass flux loss or gain during the breakup process; and, the mass flow rate at surface 1 and surface 2 should be equal on a time averaged base. That is

$$\frac{\pi}{6}\rho \sum_i \sum_j P_{i,j} D_i^3 \dot{n} = \dot{m}_l = \frac{\pi}{4}\rho D_{jet}^2 U_{jet} \qquad \text{Eq. (1)}$$

Similarly, a balance in the momentum and energy fluxes of liquid jet and droplets yield

momentum : $$\frac{\pi}{6}\rho \sum_i \sum_j P_{i,j} D_i^3 U_j \dot{n} = \dot{m}_l U_{jet} + S_{mv} = \frac{\pi}{4}\rho D_{jet}^2 U_{jet}^2 + S_{mv} \qquad \text{Eq. (2)}$$

kinetic energy : $\frac{1}{2}\frac{\pi}{6}\rho \sum_i \sum_j P_{i,j} D_i^3 U_j^2 \dot{n} = \frac{1}{2}\dot{m}_l U_{jet}^2 + S_{ke} = \frac{1}{2}\frac{\pi}{4}\rho D_{jet}^2 U_{jet}^3 + S_{ke}$

Eq. (3)

surface energy : $\pi \sigma \sum_i \sum_j P_{i,j} D_i^2 \dot{n} = \pi \sigma D U_{jet} + S_{se}$ Eq. (4)

where P_{ij} is the joint probability of finding a droplet with diameter D_i and velocity U_j; S_{mv}, S_{ke} and S_{se} are the momentum, kinetic energy and surface energy source terms respectively. They include the effects of transport between two phases (liquid & air) within the control volume. These three source terms could be positive, zero or negative. In addition, there is requirement that the sum of joint probabilities equal to unity. That is

normalization : $\sum_i \sum_j P_{i,j} = 1$ Eq. (5)

The droplet size and the droplet velocity are nondimensionalized by D_{30} (mass mean diameter) and U_{jet} (initial jet velocity) respectively. Thus, the five constraints expressed in integral form are as follows :

Normalization : $\iint f \, d\overline{D} \, d\overline{U} = 1$ Eq. (6)

Mass conservation : $\iint f \overline{D}^3 \, d\overline{D} \, d\overline{U} = 1$ Eq. (7)

Momentum conservation : $\iint f \overline{D}^3 \overline{U} \, d\overline{D} \, d\overline{U} = 1 + \overline{S}_{mv}$ Eq. (8)

Kinetic energy conservation : $\iint f \overline{D}^3 \overline{U}^2 \, d\overline{D} \, d\overline{U} = 1 + \overline{S}_{ke}$ Eq. (9)

Surface energy conservation : $B \iint f \overline{D}^2 \, d\overline{D} \, d\overline{U} = \frac{2B}{3\overline{D}_{jet}} + \overline{S}_{se}$ Eq. (10)

where f, the continuous joint probability density function (PDF) of $P_{i,j}$, can be obtained by maximizing the Shannon's entropy subject to the above five constraints. Thus, f will have the following form :

$$f = 3\,\overline{D}^2 \exp\{-\alpha_o - \alpha_1 \overline{D}^3 - \alpha_2 \overline{D}^3 \overline{U} - \alpha_3 \overline{D}^3 \overline{U}^2 - \alpha_4 B \overline{D}^2\} \qquad \text{Eq. (11)}$$

where

$$B = \frac{12}{We} \qquad We = \frac{\rho\, U_{jet}^2 D_{30}}{\sigma}$$

The formulation in the present study is slightly different from that of Li and Tankin (1989). The previous formulation conserves the combined energy flux - kinetic energy flux and surface energy flux - as one single constraint instead of two separate constraints. The shortcoming of previous formulation is that no information is provided as to how the total energy source is distributed between the kinetic energy and the surface energy. Any combination of constant total energy source will result in same probability density function. However, the kinetic energy source term primarily affects the droplet velocity distribution and the surface energy source term primarily affects the droplet size distribution. The difference between the distributions based on these two formulations will be clearer when the calculated results are presented later.

IIIb.Estimates of source terms

Due to the complexity of breakup mechanism for a spray system, normally the source terms $(\overline{S_{mv}}, \overline{S_{ke}}, \overline{S_{se}})$, can be roughly estimated according to some simple physical phenomena (e.g. Li et al., 1990). In the present study, the disentrigation of the liquid cylinder results in a spray of droplets of nearly uniform size and uniform velocity. This implies that the distributions are very steep and narrow. Under such circumstance, the results obtained from this formulation are extremely sensitive to the source terms assigned. A minor variation of a source term (less than 1%) could change the droplet distributions dramatically. This situation requires an accurate estimate of source terms for this problem to yield reliable results .

Estimating these source terms accurately will require the use of some average quantities that are obtained from experimental measurements - such

as mass mean diameter, mean velocity, etc. The relationships between source terms and experimental data are derived as follows :

$$\bar{S}_{mv} = \iint f \bar{D}^3 \bar{U} \, d\bar{D} \, d\bar{U} - 1 = \bar{D}_{30}^3 \iint f \bar{U} \, d\bar{D} \, d\bar{U} - 1 = \bar{U}_m \bar{D}_{30}^3 - 1 = \bar{U}_m - 1 \qquad \text{Eq. (12)}$$

$$\bar{S}_{ke} = \iint f \bar{D}^3 \bar{U}^2 \, d\bar{D} \, d\bar{U} - 1 = \bar{D}_{30}^3 \iint f \bar{U}^2 \, d\bar{D} \, d\bar{U} - 1$$

$$= \bar{D}_{30}^3 \iint f (\bar{U} - \bar{U}_m)^2 \, d\bar{D} \, d\bar{U} - 1 + \bar{D}_{30}^3 \iint f \bar{U}_m^2 \, d\bar{D} \, d\bar{U}$$

$$= \bar{D}_{30}^3 (\bar{U}_m^2 + \bar{U}_{rms}^2) - 1 = (\bar{U}_m^2 + \bar{U}_{rms}^2) - 1 \qquad \text{Eq. (13)}$$

$$\bar{S}_{se} = B \iint f \bar{D}^2 \, d\bar{D} \, d\bar{U} - \frac{2B}{3\bar{D}_{jet}} = \frac{B}{\bar{D}_{32}} - \frac{2B}{3\bar{D}_{jet}} \qquad \text{Eq. (14)}$$

where $\bar{D}_{30}$ is the mass mean diameter (equal to one) from experiments

$\bar{U}_m$ is the mean droplet velocity from experiments

$\bar{U}_{rms}$ is the root mean square of droplet velocity fluctuations from experiments.

$\bar{D}_{32}$ is the Sauter mean diameter from experiments

$\bar{D}_{jet}$ is the diameter of cylindrical jet at the orifice

$\frac{2B}{3\bar{D}_{jet}}$ is the surface energy flux of the cylindrical jet at the orifice.

In making these estimates of $\bar{S}_{mv}$ and $\bar{S}_{ke}$, it was assumed that all the droplets have same velocity so that $\bar{D}_{30}$ can be extracted from the integrals. That is, the velocity and size are uncoupled. This will be seen to be a reasonable assumption for the sprays under consideration.

IV.Results and Discussion

In this section, two sets of experimental data will be chosen to illustrate their corresponding distribution functions for droplet size and droplet velocity. In one set of data, the distributions are very narrow - almost delta

functions; in the other set of data, the distributions are broad. We have selected these two sets of data because they represent the two extremes from the seven sets of data collected. Figures 4 and 5 show these two sets of experimental data; one set has an acceptance ratio of 1.0 (highest) and the other an acceptance ratio of 0.91 (lowest). The narrow distributions, as in Figure 4, occur over a very restricted range of nozzle exit velocities, U_{jet}, . It was only with great care (remember these were not externally stimulated jets) that these data were obtained. The nozzle exit velocity (U_{jet}) - based on a jet diameter of 50 μm - equals 2.76m/sec in Figure 4. Changing U_{jet} (either increasing or decreasing) results in broader distributions. For example, the distributions at U_{jet} = 2.32 m/sec is shown in Figure 5. Similar results (but not quite as broad as those in Figure 5) occurred when U_{jet} = 3.64 m/sec. Intermediate widths of the distributions occur when 3.64m/sec > U_{jet} > 2.76 m/sec and 2.32m/sec< U_{jet} < 2.76 m/sec.

For each set of experiments there are two distribution curves - one is number of droplets vs. droplet diameter (4a and 5a) and the other one is number of droplets vs. droplet velocity (4b and 5b). Accompanying these two plots are the experimental parameters and measured mean values of droplets. $\overline{D}_{30}$, $\overline{D}_{32}$, $\overline{U}_m$, and $\overline{U}_{rms}$ are used in estimating the source terms. It was assumed in these initial estimates (Equations 12, 13 and 14) that the droplet size is independent of droplet velocity. This assumption needs justification.

The assumption for estimating $\overline{S}_{mv}$ and $\overline{S}_{ke}$ can be verified from the experimental data recorded by the PDPA. If one examines the raw data - for example, for the test shown in Figure 5b - one can plot the mean diameter ($\overline{D}_{30}$) and its RMS value for each velocity bin. This is shown in Figure 6a for velocity bins having more than 10 drops/bin. The mean diameter is nearly independent of velocity and the RMS of the drop diameter in each velocity bin is small compared to the mean diameter (1). The largest deviation of $\overline{D}_{30}$ from 1 occurs at velocity bins (U/Ujet) greater than 0.35; but even in these bins the D/D_{30} is greater than 0.97 and the RMS is less than 0.15. In addition, the number of drops in the velocity bins greater than 0.35 only amounts to 10% of the total drops measured. Thus the assumption that the droplet

diameter is independent of velocity is valid. Therefore one would expect the estimated source terms to be reasonably accurate.

As an alternate approach, one could assume a uniform velocity ($\overline{U}_m$) instead of a unifrom droplet size when making estimates for $\overline{S}_{mv}$ and $\overline{S}_{ke}$ in Equations 12 and 13. By employing this assumption, the only difference from the previous approach is that $\overline{S_{ke}}=\overline{U_m^2}-1$. The term $\overline{U_{rms}^2}$ disappears due to the assumption that the velocity is constant. It is interesting to note that the influence of the droplet size distribution on the estimates of the source terms is the same in these two appraches. The assumption of uniform velocity is verfied experimentally by plotting the mean and RMS values of the droplet velocity in each size bin as shown in Figure 6b. The RMS values for the velocity in each bin is less than 0.02 (6%).

A third approach can be used without making any assumption about the droplet size or droplet velocity distributions. This approach is based two additional mean variables which are not directly measured by PDPA. If one defines the mean velocity and RMS values of the velocity in terms of droplet volume instead of droplet number, then the source terms $\overline{S}_{mv}$ and $\overline{S}_{ke}$ can be derived by simply replacing the number-based mean velocity and the RMS of the velocity in Equations 12 and 13 with volume-based variables. Although these two mean variables are not directly available from the standard PDPA output, they could be obtained from the raw data.

The droplet distributions obtained from these estimated source terms only provide an initial result, which will probably require modification of the source terms in order to yield distributions that agree with the experimental data. Therefore we will not concern ourselves with the different possible results obtained from these three approaches for estimating the source terms, but choose the first approach - constant droplet diameter. It is believed that any one of these three approaches will provide sufficiently accurate estimates of the source terms needed to make the initial calculations. Once the source terms are estimated, the corresponding distribution functions of droplet size

and droplet velocity are obtained by integrating Equation (11) along $\overline{D}$ and $\overline{U}$ respectively. They are :

$$\frac{dN}{d\overline{D}} = \frac{3}{2}\left(\frac{\pi \overline{D}}{\alpha_3}\right)^{\frac{1}{2}} [\, \mathrm{erf}(X_{max}) - \mathrm{erf}(X_{min}) \,] \exp\{-\alpha_o - (\alpha_1 - \frac{\alpha_2^2}{4\alpha_3})\overline{D}^3 - \alpha_4 B \overline{D}^2\} \quad \text{Eq. (15)}$$

$$\frac{dN}{d\overline{U}} = \int f \, d\overline{D} \quad \text{Eq. (16)}$$

where erf (X) is the error function of X

$$X_{max} = (\overline{U}_{max} + \frac{\alpha_2}{2\alpha_3})(\alpha_3 \overline{D}^3)^{\frac{1}{2}}$$

$$X_{min} = (\overline{U}_{min} + \frac{\alpha_2}{2\alpha_3})(\alpha_3 \overline{D}^3)^{\frac{1}{2}}$$

Since there is no closed form solution for the droplet velocity distribution, Equation (16) is calculated by numerical integration. N is the normalized droplet number. "Estimated" distributions (based on estimated source terms) for these two set of data are also shown in Figures 7 and 8. The estimated source terms and the corresponding parameters (α_o, α_1, α_2, α_3, α_4) for these predicted distributions are listed in Tables I & II. Agreement between measurements and estimated correlations is reasonably good. However, there is a discrepancy between the peaks of estimated distributions and the experimental data.

The "estimated" distribution function was further modified by adjusting the source terms - resulting in "fitted" distribution functions. The fitting technique applied, in this case, relies on the fact that the We number and the surface energy source term $\overline{S}_{se}$ have a major effect on the droplet size distribution and a minor effect on the velocity distribution. On the other hand, the other two source terms $\overline{S}_{mv}$ and $\overline{S}_{ke}$ have a major effect on the velocity distribution, but minor effect on the size distribution. The procedure used for fitting is described briefly as follows :

(1). To fit the droplet size distribution, the We number was fixed and the surface energy source $\overline{S}_{se}$ was adjust until the maximum amplitude of

$dN/d\overline{D}$ of the experimental data and the fitted data matched. In this manner, the size distribution is fitted while the velocity distribution remains essentially unchanged.

(2). To fit the droplet velocity distribution, the source term $\overline{S}_{mv}$ was fixed and the kinetic energy source term $\overline{S}_{ke}$ was adjusted. Thus the velocity distribution is fitted while the size distribution remains essentially unchanged.

The results for the combined changes in $\overline{S}_{se}$ and $\overline{S}_{ke}$ are shown in Figures 7 and 8. The corresponding parameters and source terms are listed in Tables I and II. The difference in the source terms used in the "estimated" and "fitted" curves varied by only a fraction of a percent! For example, in Figure 7, the difference in $\overline{S}_{se}$ is only 0.0038% ; the difference in $\overline{S}_{ke}$ is only 0.00016%. In Figure 8 these differences are only 0.0001% and 0.259%. Similar results were obtained for the other five sets of data collected and examined. Thus, one can see the importance of "accurately" estimating the source terms.

The distributions using the previous formulation (Li & Tankin, 1989) of a single energy flux constraint was examined. Attempts were made to determine the corresponding droplet distributions for the two sets of data based on the formulation that there is only one constraint for the energy flux. The combined energy source term is the sum of two separate energy source terms ($\overline{S}_{se}$ and $\overline{S}_{ke}$) estimated earlier (see Tables I & II). For data set in Figure 7, the numerical calculations didn't converge. The results for data set in Figure 8, reveal that the velocity and size distributions disagree with the experimental data. Thus the conservation of combined energy source term doesn't provide enough information for this spray system.

Finally the computed flow rate using the PDPA measurements is compared with the measured flow rate obtained by weighing a sample of the droplets collected over a period of time. The following equation was used to compute the flow rate:

$$\dot{Q} = \pi \sum_i \dot{n}_i \frac{D_i^3}{6} \qquad \text{Eq. (17)}$$

where

D_i is the diameter of the droplet in the "i"th bin

$\dot{n}_i$ is the number of droplets per unit time in the "i"th bin.

A comparison of the difference between the calculated flow rate using the PDPA and the measured flow rate is 8%. However, since the diameter is raised to the third power, the measured diameter of the droplets is within 3% of the actual diameter. This implies a consistency between the measured data and the PDPA data. In making the calculation represented by Equation 17, the bin diameter, (D_i), is the lower limit of the ith bin. If one used an average bin diameter of the ith bin, the difference between the metered flow rate and the calculated flow rate would be about 3%; thus resulting in a discrepancy of only 1% in droplet diameter.

V. Conclusions

The agreement between theory and experiments is reasonably good for the Rayleigh problem under investigation. In order to obtain these results it was necessary to modify the previous theory of Li and Tankin (1987, 1988, 1989), by separating the energy source term into a kinetic energy source term and a surface energy source term. It was possible for the breakup of a cylindrical jet to evaluate the source terms - within a fraction of a percent. In the experiments conducted in this study it was not necessary to involve the probe area of the PDPA (a possible source of error) since all the droplets passed through the measuring volume. Also the acceptance ratio exceeded 0.91 in these experiments. The measured liquid flow rate agrees to within 8% that obtained from the PDPA measurements.

VI. Acknowledgements

The authors wish to thank the assistance provided by Fluid Jet Associates and in particular Dr. J. L. Dressler, president. He supplied the nozzle and supervised the operation of the droplet generator.

VII. References :

Bogy, D. B., Drop Formation in a Circular Liquid Jet, *Ann. Rev. Fluid Mech.*, Vol. 11, 1979.

Bousfield, D. W., Stockel, I. H. and Nanivadekar, C. K., The Breakup of Viscous Jets with Large Velocity Modulations, *J. Fluid Mech.*, Vol. 218, 1990

Chaudhary, K. C. and Maxworthy, T., The Nonlinear Capillary Instability of a Liquid Jet, Part 3 : Experiments on Satellite Drop Formation and Control, *J. Fluid Mech.* Vol. 96, 1980.

Dressler, J. L. and Kraemer, G. O., A Multiple Drop-Size Drop Generator for Calibration of a Phase - Doppler Particle Analyzer, in *Liquid Particle Size Measurement Techniques - Vol. II*, 1990.

Goedde, E. F., and Yuen, M. C., Experiments on Liquid Jet Instability, *J. Fluid Mech.* Vol. 40, 1970.

Haken, H., *An Introduction to Synergetics*. Springer-Verlag, Berlin, Heidelberg, 1978.

Jaynes, E. T., Information theory and statistical mechanics. *Phy. Rev.* Vol. 106; Vol. 108, 1957.

Kelly, A. J, Electrostatic metallic spray theory. *J. of Applied Physics*, Vol. 47, 1976.

Li, Xianguo and Tankin, R. S., Droplet Size Distribution: A Derivation of a Nukiyama - Tanasawa Type Distribution Function, *Combust. Sci. and Tech.*, Vol. 56, 1987.

Li, Xianguo and Tankin, R. S., Derivation of Droplet Size Distribution in Sprays by Using Information Theory, *Combust. Sci. and Tech.* , Vol. 60, 1988.

Li, Xianguo and Tankin, R. S., Prediction of Droplet Size and Velocity Distributions in Sprays Using Maximum Entropy Principle, *Combust. Sci. and Tech.*, Vol. 68, 1989.

Li, X., Chin, L. P., Tankin, R. S., Jackson, T., Stutrud, J. and Switzer, G., Comparison Between Experiments and Predictions Based on Maximum Entropy for Sprays from a Pressure Atomizer, *Combustion and Flame*, 1990, to be published.

Pimbley, W. T. and Lee, H. C., Satellite Droplet Formation in a Liquid Jet, *IBM J. Res. Dev.* Vol. 21, 1977.

Rayleigh, L. On the Instability of Jets. *Proc. London Math. Soc.*, Vol. 10, 1879.

Rayleigh, L. Further Observations upon Liquid Jets, *Proc. R. Soc. London*, Vol. 34, 1882.

Sellens, R. W. and Brzustowski, T. A., A Prediction of the Drop Size Distribution in a Spray from First Principle. *Atomization and Spray Tech.* Vol. 1, 1985.

Sellens, R. W. and Brzustowski, T. A., A Simplified Prediction of Droplet Velocity Distributions in a Spray. *Combust. and Flame* Vol. 65, 1986

Shannon, C. E., A Mathematical Theory of Communication. *The Bell System Technical Journal* Vol. 27, 1948, Also in D. Slepiom (Ed.), *Key Papers in the Development of Information Theory*, IEEE press, New York, 1974.

Tribus, M., *Thermostatics and Thermodynamics*. D. Van Nostrand, N. Y., 1961.

VIII. Figure Captions

Figure 1. Schematic drawing of the droplet generator setup.

Figure 2. Photograph showing the stream of droplets produced by the droplet generator.

Figure 3. Control volume used in this study for the liquid jet and droplets .

Figure 4. Print-out of the measurements obtained from PDPA for data set #1. Figure 4a is the number of droplets versus droplet diameter, and Figure 4b is the number of droplets versus droplet velocity.

Figure 5. Print-out of the measurements obtained from PDPA for data set #2. Figure 5a is the number of droplets versus droplet diameter, and Figure 5b is the number of droplets versus droplet velocity.

Figure 6. Plots showing the dependence of the droplet velocity on droplet diameter. Figure 6a is a plot of the mean velocity and its rms values versus droplet diameter. Figure 6b is a plot of the mean diameter and its rms values versus droplet velocity.

Figure 7. Comparison between experimental data ( ), "estimated", and "fitted" results for data set # 1. In Figure 7a are the droplet size distributions and in Figure 7b are the droplet velocity distributions.

Figure 8. Comparison between experimental data ( ), "estimated", and "fitted" results for data set # 2. In Figure 8a are the droplet size distributions and in Figure 8b are the droplet velocity distributions.

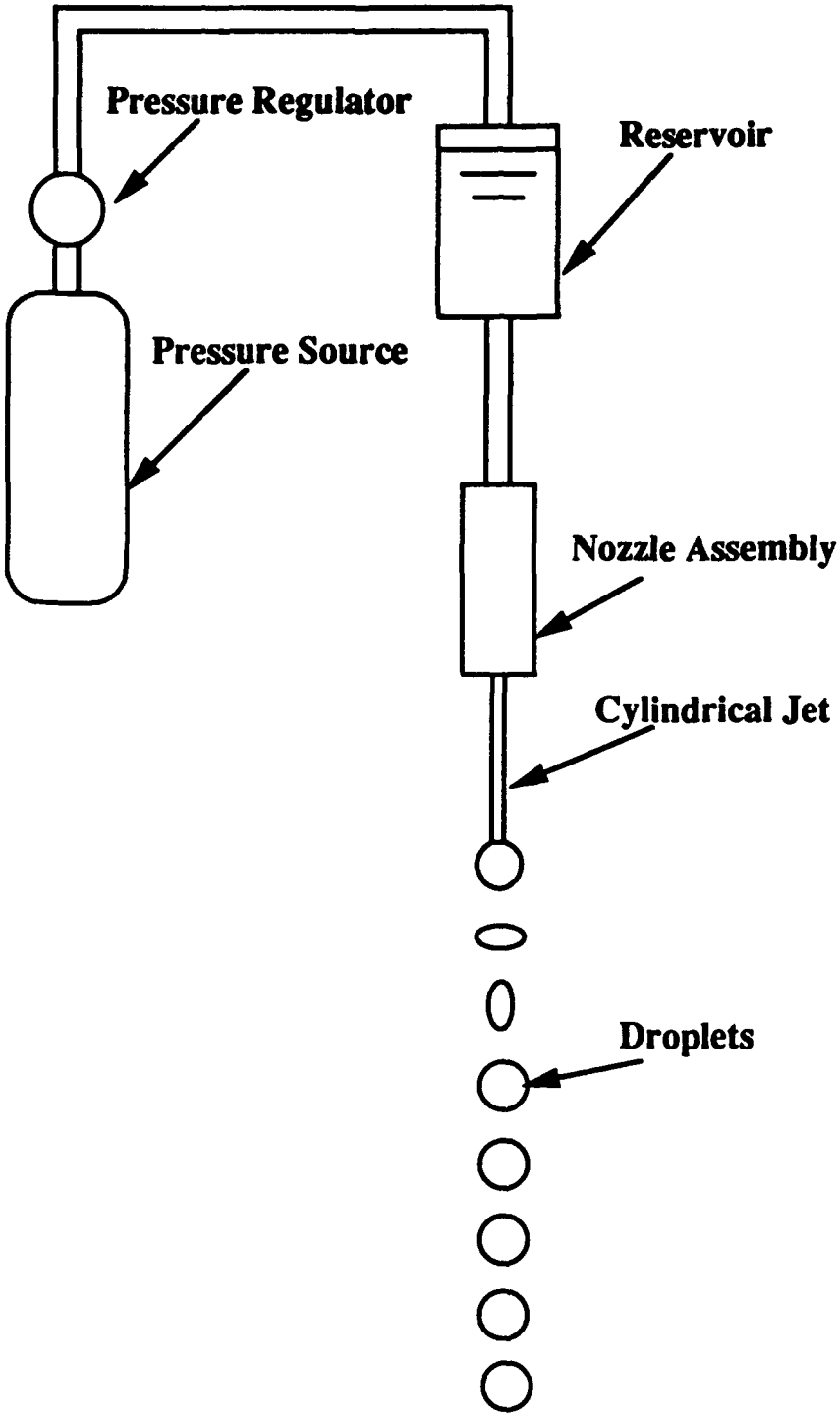


Figure 1 Schematic drawing of droplet generator

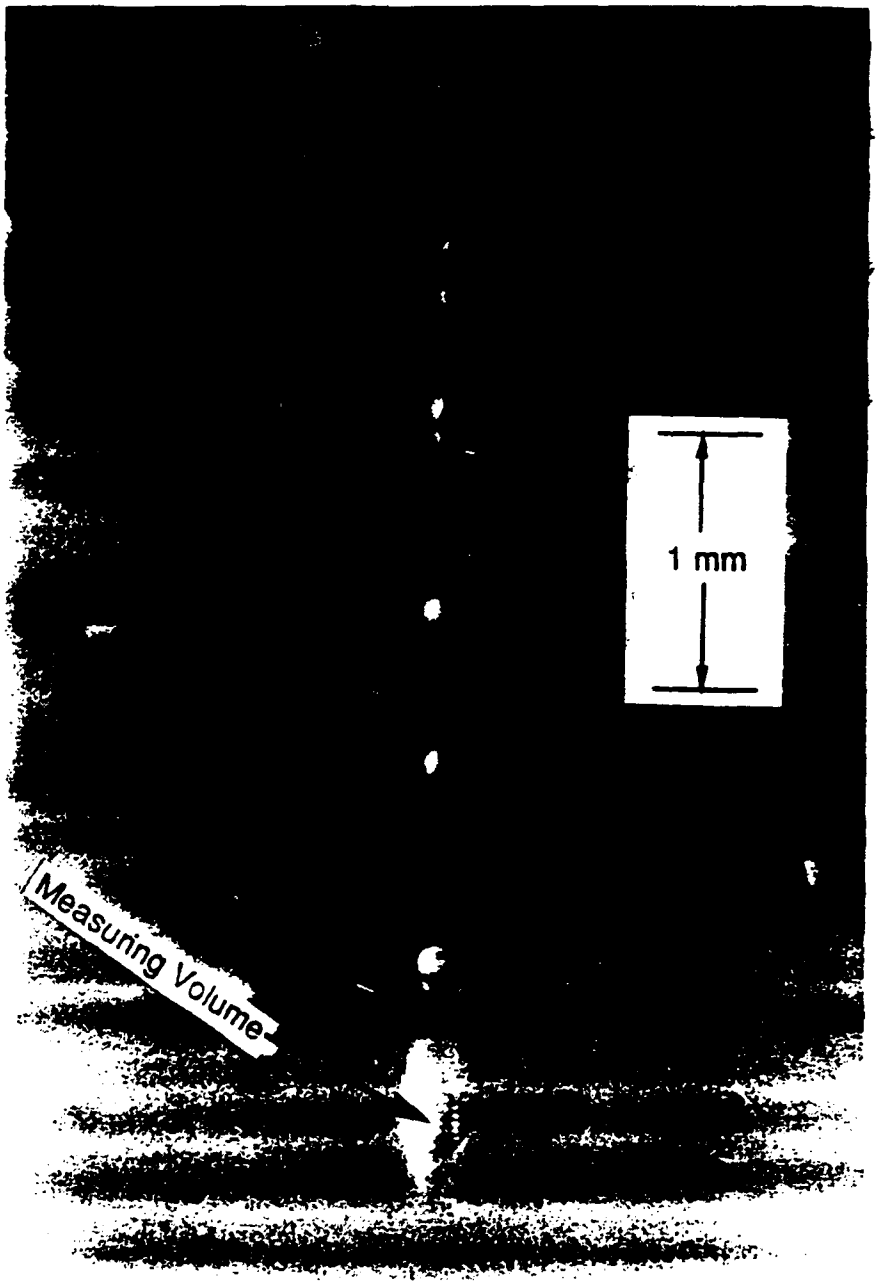


FIG. 2. Photograph showing the stream of droplets produced by the droplet generator

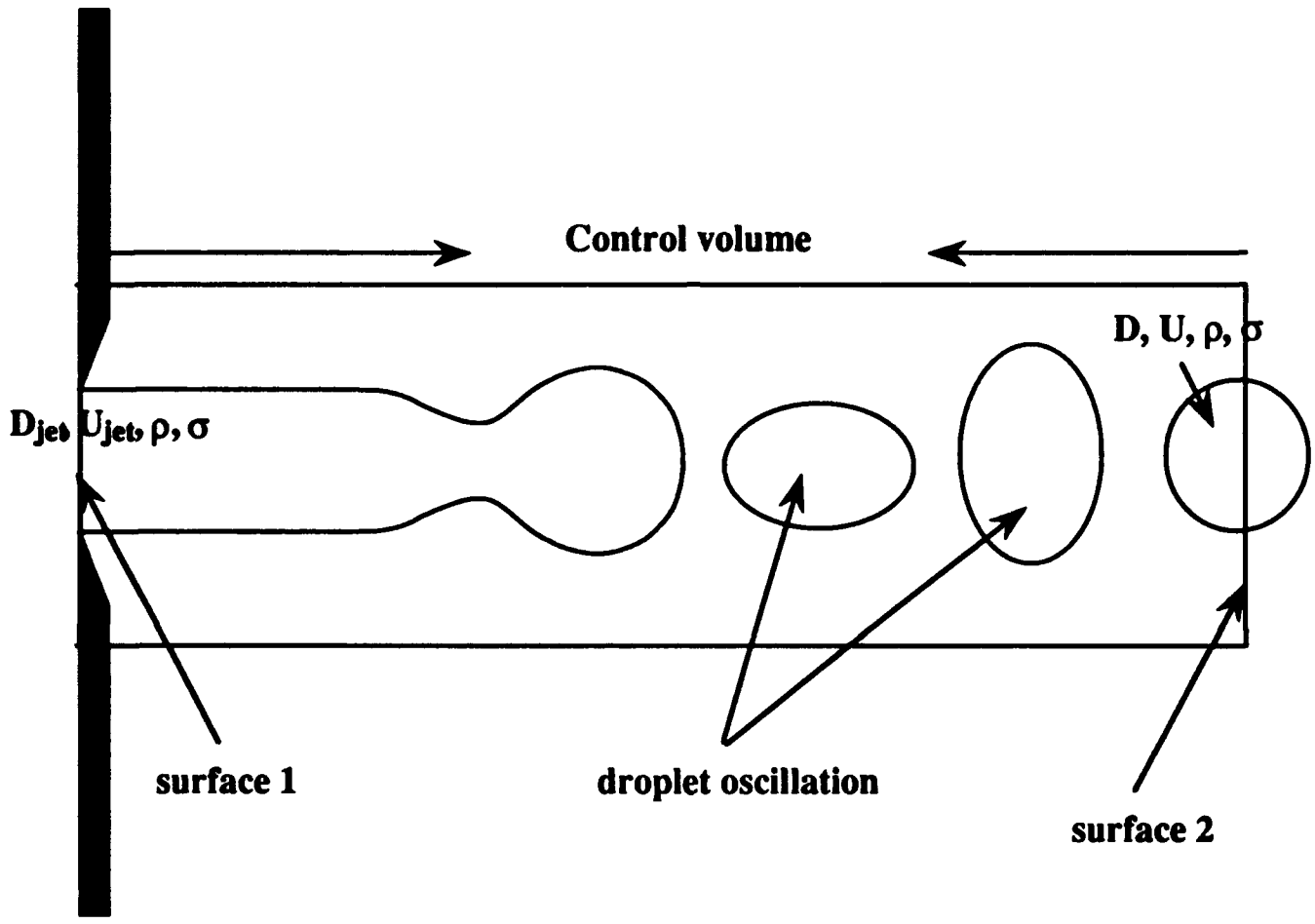
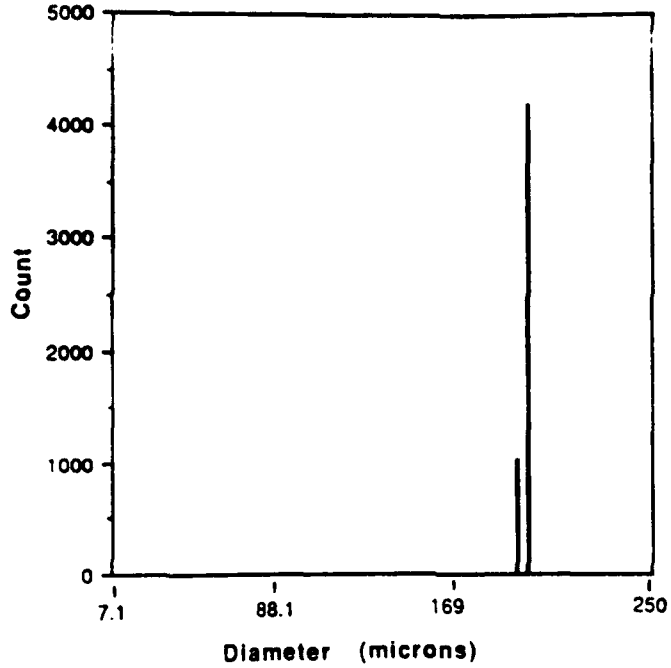
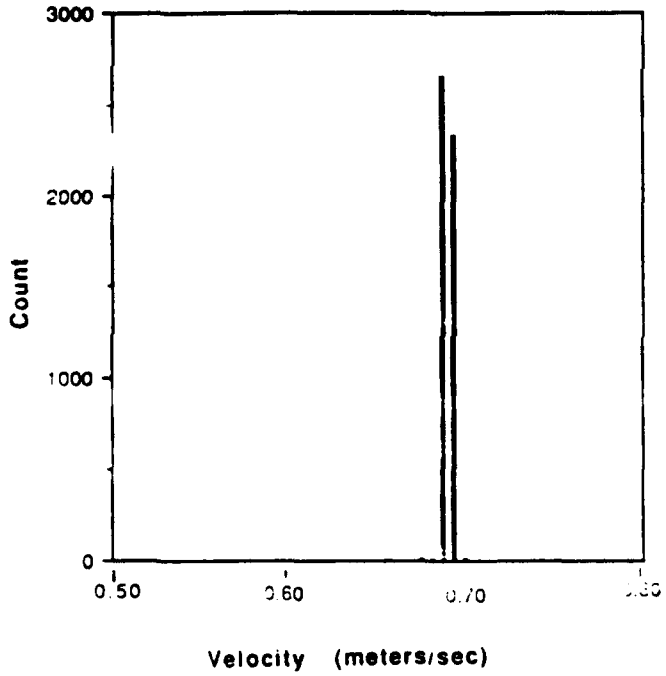


Figure 3 Liquid jet, formed from an orifice, breaking into droplets in the Control volume.



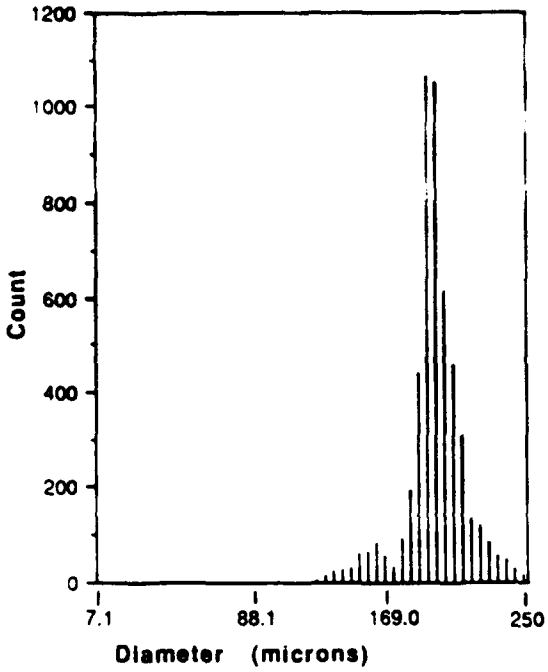
(a)



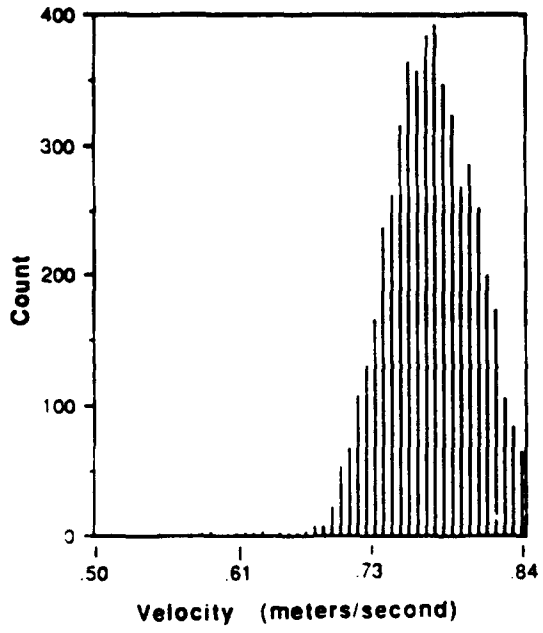
(b)

Arithmetic Mean	(D_{10}) = 194.5 μm
Area Mean	(D_{20}) = 194.5 μm
Volume Mean	(D_{30}) = 194.5 μm
Sauter Mean	(D_{32}) = 194.6 μm
Probe Area	= 9.92 E-004 cm^2
Number Density	= 2.07 E+004 /cc
Vol. Flow Rate	= 5.43 E-003 cc/s
Volume Flux	= 5.47 $cc/s/cm^2$
Attempts	= 5022
Validations	= 5022
Corrected Count	= 5210
Run Time	= 3.57 sec
Mean Velocity	= 0.687 m/s
RMS Velocity	= 0.003 m/s

FIG. 4 Printout of the measurements obtained from PDPA for data set No. 1. (a) The number of droplets versus droplet diameter and (b) the number of droplets versus droplet velocity.



(a)



(b)

Arithmetic Mean (D_{10}) = 199.7 μm
Area Mean (D_{20}) = 200.4 μm
Volume Mean (D_{30}) = 201.1 μm
Sauter Mean (D_{32}) = 202.4 μm

Probe Area = 1.62 E-003 cm^2
Number Density = 8.66 E+003 /cc
Vol. Flow Rate = 4.56 $cc/s/cm^2$

Attempts = 5492
Validations = 5004
Corrected Count = 5148
Run Time = 4.00 sec

Mean Velocity = 0.766 m/s
RMS Velocity = 0.035 m/s

FIG. 5 Printout of the measurements obtained from PDPA for data set No. 2 (a) The number of droplets versus droplet diameter and (b) the num [illegible]

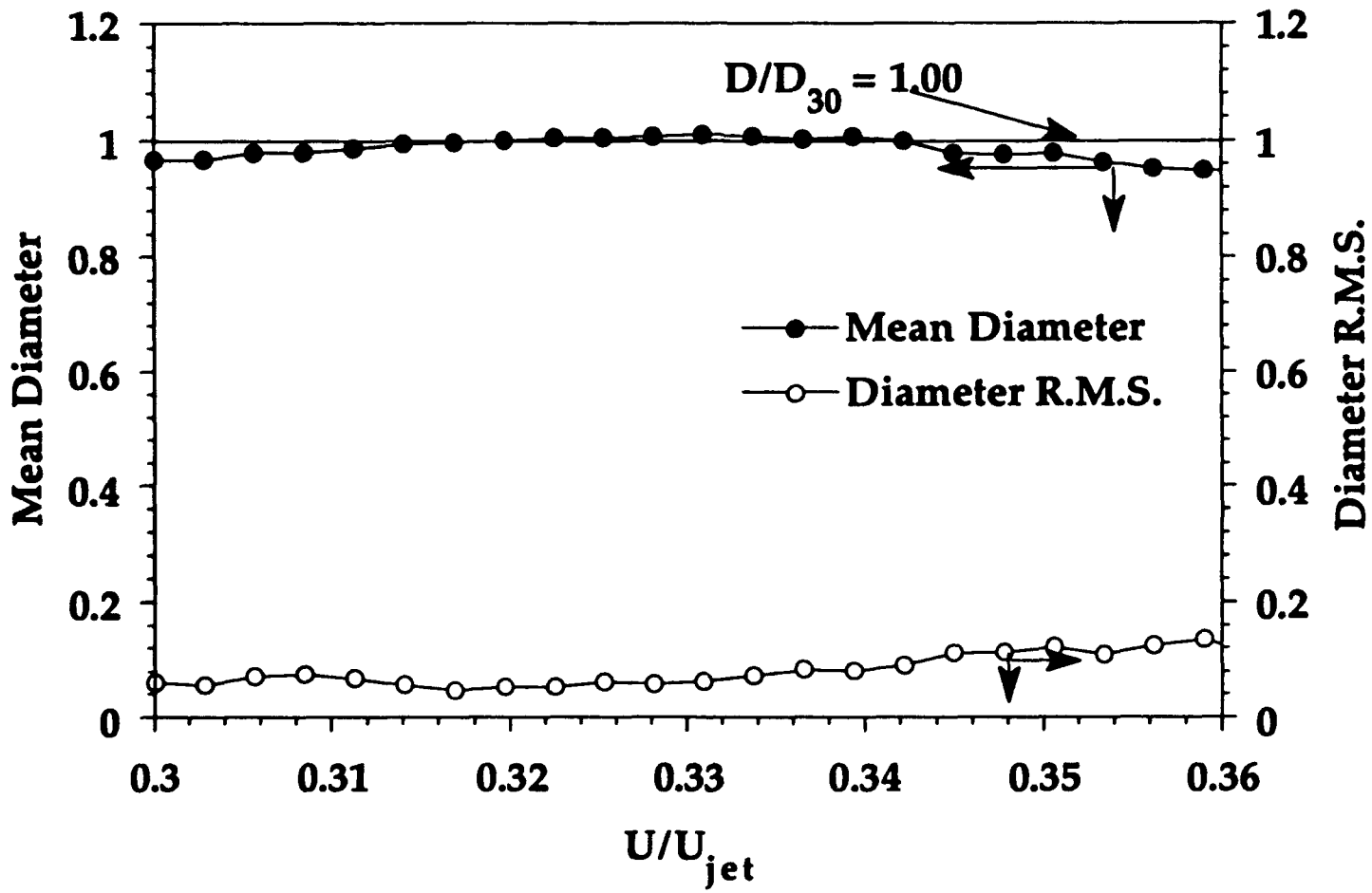


Figure 6(a) Mean diameter and diameter RMS vs. droplet velocity

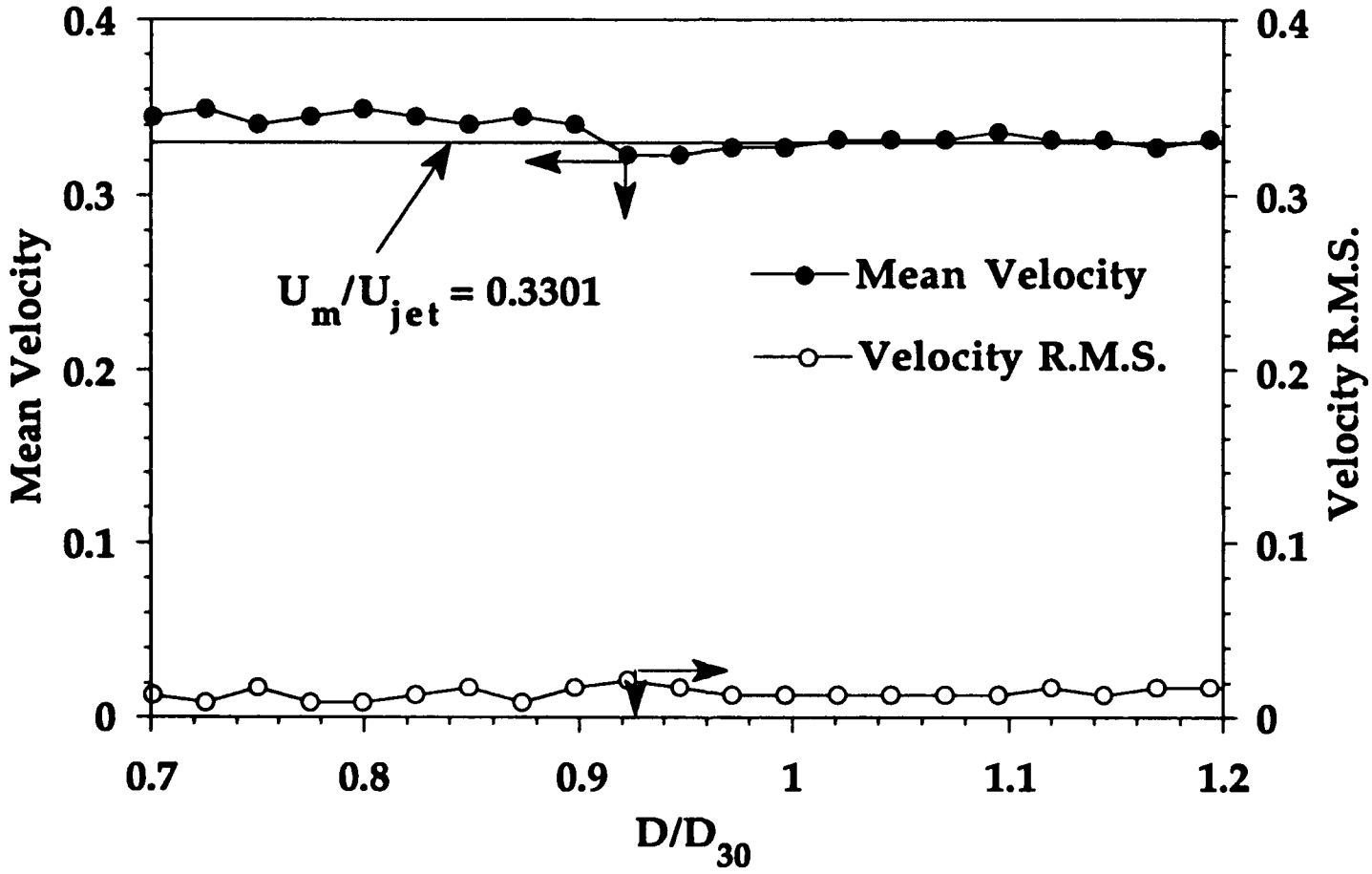


Figure 6(b) Mean velocity and velocity RMS vs. droplet diameter

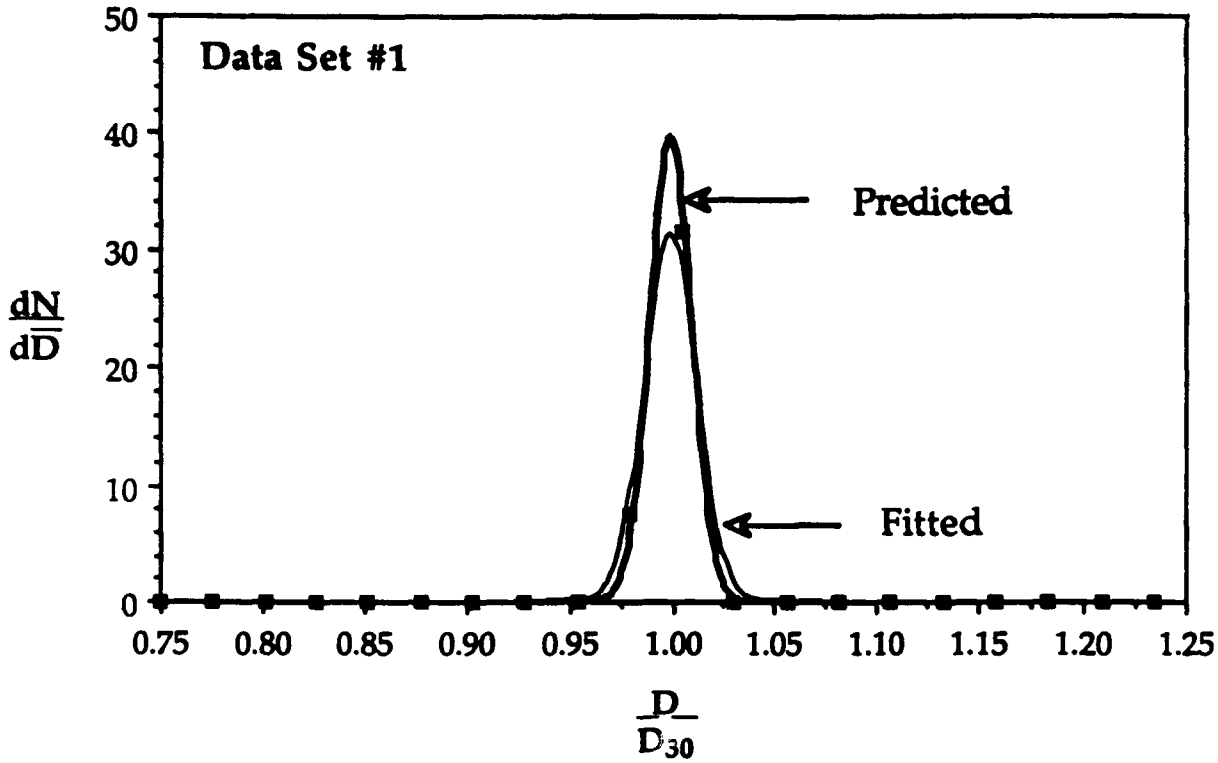


Figure 7(a) Droplet size distributions for data set #1

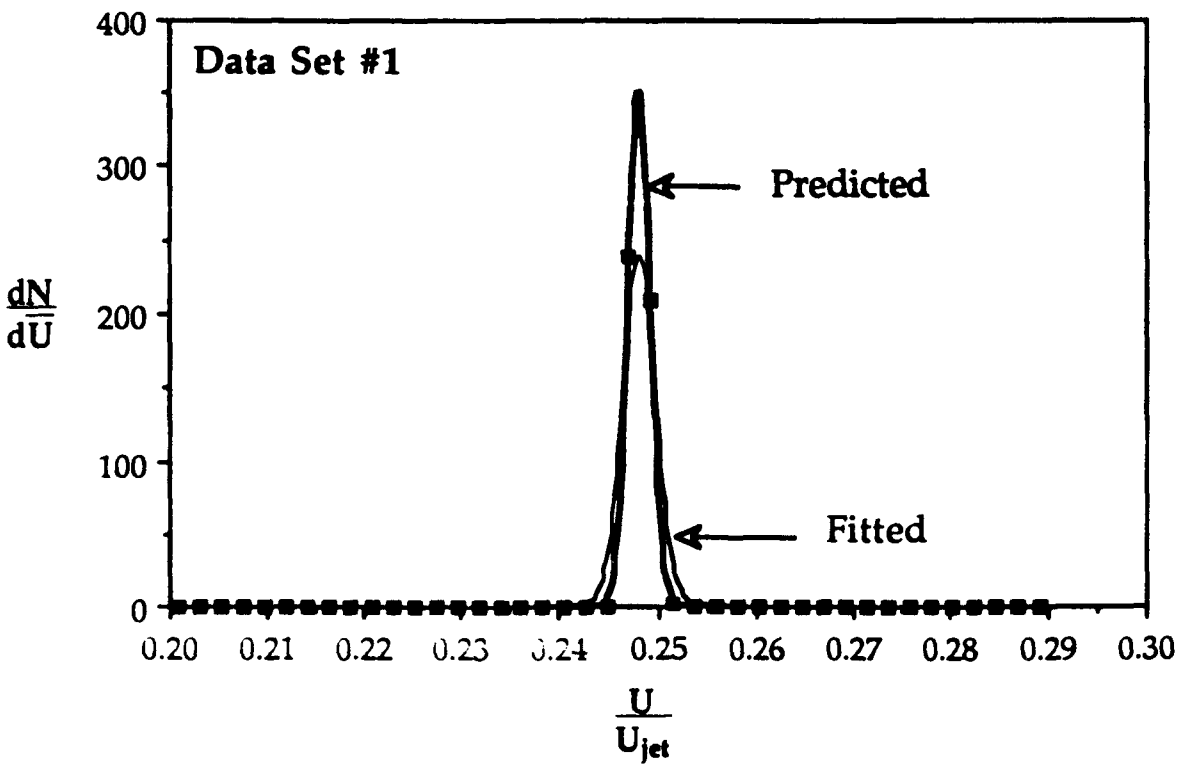


Figure 7(b) Droplet velocity distributions for data set #1

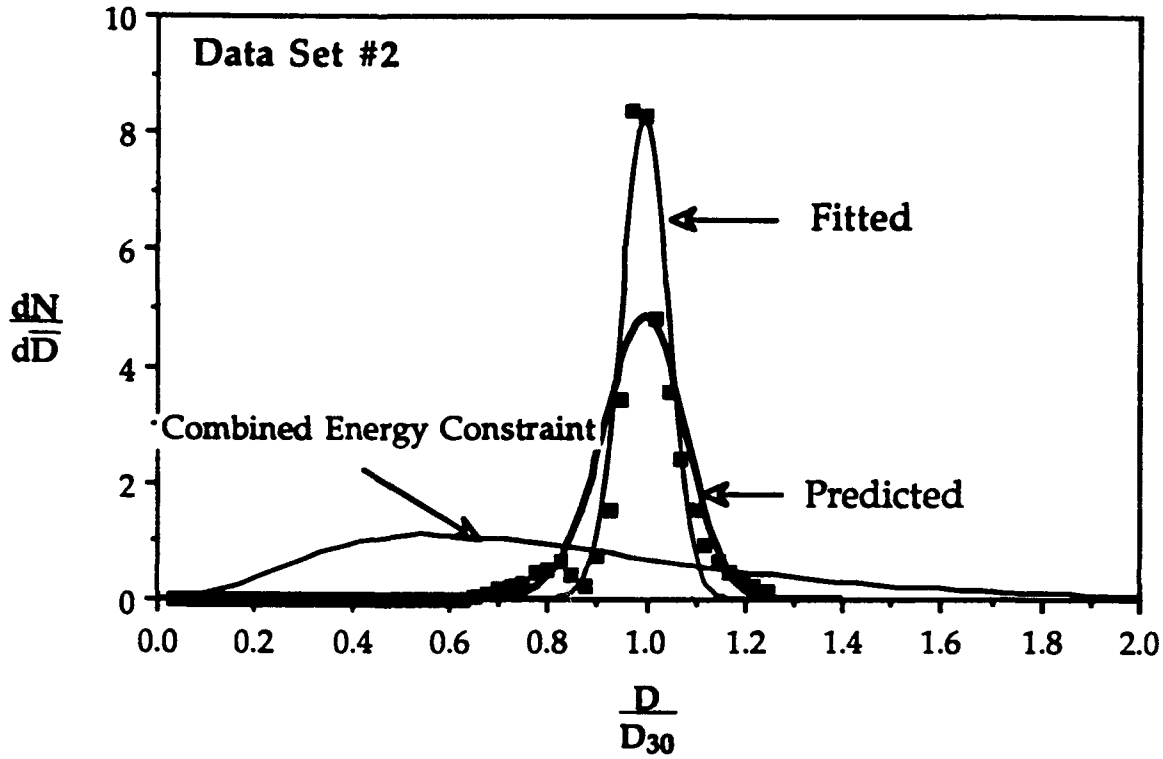


Figure 8(a) Droplet size distributions for data set #2

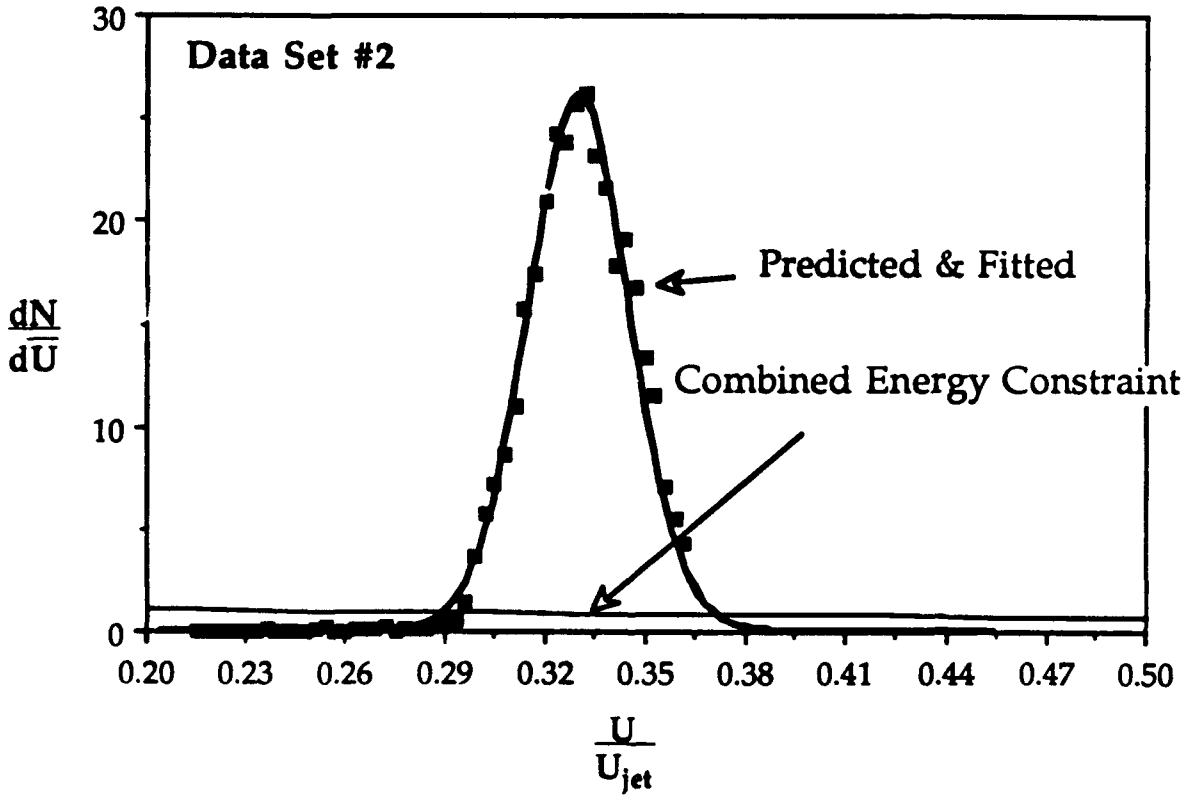


Figure 8(b) Droplet velocity distributions for data set #2

	correlation	fitting
We	20.40002	20.40002
$\bar{S}_{mv}$	-0.7517503	-0.7517503
$\bar{S}_{ke}$	-0.9383708	-0.9383693
$\bar{S}_{se}$	-0.9376119	-0.9376480
α_0	1639.5249	1017.7057
α_1	27249.1459	13110.0542
α_2	-192972.6902	-89091.7504
α_3	388666.5124	179439.7947
α_4	-8405.5046	-5231.04300

Table I Source terms and distribution parameters for data set #1

	correlation	fitting
We	14.87016	14.87016
$\bar{S}_{mv}$	-0.6698716	-0.6698716
$\bar{S}_{ke}$	-0.8907844	-0.8907854
$\bar{S}_{se}$	-1.3619760	-1.3584500
α_0	20.9063	66.3845
α_1	285.8760	378.9142
α_2	-1430.1208	-1436.3431
α_3	2166.0070	2175.4309
α_4	-92.2844	-263.3113

Table II Source terms and distribution parameters for data set #2

1991 USAF-UES Faculty Research Initiation Program

Follow-Up Grant

Sponsored by the

Air Force Office of Scientific Research

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT:
FUNCTION DECOMPOSITION AND MACHINE–LEARNING SYSTEMS

Mike Breen

Mathematics Department

Alfred University

Alfred, NY 14802

December 4, 1991

Function Decomposition and Machine-Learning Systems

by

Mike Breen

ABSTRACT

The concept of a function is central to the mission of fire control. We want the hardware to learn the desired function – to give the desired output when presented with input. With this in mind, efficiency can be best achieved by finding the simplest form of the desired function. Several experiments have been done that suggest that learning is best done with the technique known as Function Extrapolation by Recomposing Decompositions (FERD). We present some results about FERD.

ACKNOWLEDGEMENTS

I am extremely grateful to the Air Force Office of Scientific Research, the Air Force Systems Command, and the Universal Energy Systems, Inc. for the opportunity to do the research and for their sponsorship. I would also like to thank Dr. Tim Ross, who has shown the way on this subject. Also John Jacobs, Paul Johnson, and Lt. Tim Taylor at Wright-Patterson Air Force Base have all been very helpful.

INTRODUCTION

In the field of avionics, complicated tasks must be performed quickly and efficiently. A vast amount of information must be processed while that information is still relevant. A natural conclusion is that this information-processing and decision-making can best be done with the best machine. However, with all other things being equal, simplifying the algorithm that describes the task is far more efficient than boosting the hardware.

With that in mind, the Systems Concept Group has been studying Pattern-Based Machine Learning, especially the idea of function decomposition (the process of receiving information and giving a course of action is the process of a function -- input is given to the machine and output is returned). Since there is no way to program all the possible inputs that a machine may receive, we want the machine to learn its intended function. What is the best way for the machine to learn?

Several experiments run by Dr. Ross and Lt. Taylor at WPAFB suggest that FERD offers the best approach to learning. Let us briefly explain how FERD works.

We have a function that we would like the machine to learn. Our functions are of the form $f : \{0, 1, \ldots, 2^n - 1\} \rightarrow \{0, 1\}$. We may think of f as a collection of ordered pairs $f = \{(a, b) : a \in \{0, 1, \ldots, 2^n - 1\}, b \in \{0, 1\}, f(a) = b\}$. We randomly choose a sample from f (in its set form), then find the function of least complexity that agrees with f on this sample. That function is the learned function that FERD selects.

The results obtained so far indicate that FERD generally gets the function right. That is, its learned function is the desired function (when the size of the sample is on the order of the complexity of the desired function). FERD performs better than other learning techniques that have been tried, neural nets for example.

We present here some theoretical results about FERD in the hope of trying to determine why FERD works, if it will always work in the manner described above, and the implications of this for machine learning.

OBJECTIVES

Our primary objective is to investigate function decomposition and its applicability to machine-learning. All the experiments performed to this date suggest that for the functions we are concerned with, function decomposition provides the best learning algorithm. We are investigating why this is so by formulating some of its general properties and investigating how these properties relate to specific cases.

We have also read many articles in an attempt to find where the function decomposition algorithm approach to learning fits into other areas of complexity theory and learning theory.

THE REPORT

Let us first explain our general situation. We have a function $f : \{0, 1, \ldots, 2^n - 1\} \rightarrow \{0, 1\}$ often called a binary or Boolean function. The function f can be visualized as a set of ordered pairs (a, b) where $a \in \{0, 1, \ldots, 2^n - 1\}, b \in \{0, 1\}$ and $f(a) = b$. This is the function that we want the machine to learn. To accomplish this, we select a sample of ordered pairs from f, and call this set A. Thus, $A \subseteq f$.

When the machine–learning system is given A, it then attempts to reconstruct a function, g, from this sample. Perhaps g will equal f, perhaps not. The latter case is more likely. In fact, with some machine–learning systems, the learned function, g, may not even equal f on the sampled set A ! We call the number of values at which f and g disagree -- $|\{x : f(x) \neq g(x)\}|$ — the error e. Ideally, we would like ϵ to be zero. This would mean that $g = f$ and the machine has learned the function. Note that we can always do this by letting $A = f$ in which case no learning is necessary. However, as might be expected, it is impractical if not impossible for A to equal f. For one thing, f is too large. Its ordered pair listing can contain billions of ordered pairs. For another, often we don't know all of f. In practice, $|A|$ — the number of elements in A — is far less than $|f|$.

The best learning that can be done is done by the algorithm that minimizes ϵ on the smallest A when compared to all other learning algorithms. There is not one algorithm that can do this for every function possible and every sample that we might take. However, the Function Extrapolation by Recomposing Decompositions (FERD) has consistently outperformed other algorithms when presented with many different types of binary functions. Several results on FERD are found in [4].

Let us briefly explain how it works. When the random sample, A, has been selected, FERD constructs functions with the same domain and range as f that are equal to f on A and have complexity of at most the complexity of f. Our complexity here is the

Decomposed Function Cardinality (DFC) complexity defined by Ross in [4]. We write the cost of a function θ as $C(\theta)$.

Let $F = \{g : g : \{0, 1, ..., 2^n - 1\} \rightarrow \{0, 1\}\}, S = \{g \in F : g(x) = f(x)\ \forall x \in A\}$, and $C = \{g \in F : C(g) \leq C(f)\}$. Then FERD finds a function in $S \cap C$. It finds a function of lowest cost that agrees with f on A. We call this function the learned function. Of course, the best result is when the learned function is f. In that case, FERD "gets it right".

We now present some facts about costs. Let f be given and let $A \subseteq f$. Define $\overline{C}_A(f)$ =the largest cost of the functions in S and $\underline{C}_A(f)$ =the smallest cost of the functions in S. To be precise, in the first definition, there may not be a largest cost, but what we mean is the cost such that no function in the set has a higher cost. The same is true for the second definition.

If V is a finite set of real numbers, let $min(V)$ denote the number m such that $m \leq v$ for all $v \in V$, and let $max(V)$ denote the number M such that $M \geq v$ for all $v \in V$. Recall that for any finite sets of real numbers, T and U, if $T \subseteq U$, then $min(T) \geq min(U)$ and $max(T) \leq max(U)$.

LEMMA 1. *If $A \subseteq A' \subseteq f$ then*

(1) $\underline{C}_A(f) \leq \underline{C}_{A'}(f)$; *and*

(2) $\overline{C}_A(f) \geq \overline{C}_{A'}(f)$.

PROOF: Note that if $A \subseteq A'$, then $S_A = \{g \in F : g(x) = f(x)\ \forall x \in A\} \supseteq \{g \in F : g(x) = f(x)\ \forall x \in A'\} = S_{A'}$. Both statements now follow from the above remarks since the quantities involved are the minimum and maximum, respectively. ∎

The Lemma allows us to state and prove the upcoming Theorem. It says that regardless of the sample sizes involved, the smallest costing function resulting from a sample can be no larger than the largest costing function resulting from any other sample.

THEOREM 2. *For any $f \in F$, and any $A, B \subseteq f, \underline{C}_A(f) \leq \overline{C}_B(f)$.*

PROOF: Consider $A \cup B$ which is also contained in f. Since $A, B \subseteq A \cup B$ we have by definition and by the Lemma that

$$\underline{C}_A(f) \leq \underline{C}_{A \cup B}(f) \leq \overline{C}_{A \cup B}(f) \leq \overline{C}_B(f). \quad \blacksquare$$

This allows us to visualize FERD's learned function eventually converging to f.

We now list a result similar to those above.

The learned function is obtained by chance. A random sample of f is chosen and FERD constructs a function consistent with f on that sample. We would like the probability that the learned function is f to be one but this is not usually the case. We strive to know how much of FERD is random and how much is determinable.

Probabilistically, let f be given, let $A \subseteq f$, and let $P_A(f)$ be the probability that given sample A, FERD correctly identifies f.

PROPOSITION 3. *If $A_1 \subset A_2 \subseteq f$, then $P_{A_1}(f) \leq P_{A_2}(f)$.*

PROOF: Let $A_1 \subset A_2, S_1 = \{g \in F : g(x) = f(x)\ \forall x \in A_1\}, S_2 = \{g \in F : g(x) = f(x)\ \forall x \in A_2\}$. Then, as in the Lemma, $S_2 \subseteq S_1$ so that $S_2 \cap C \subseteq S_1 \cap C$. Hence, FERD has a larger field to search through to correctly learn f when learning on A_1 as opposed to when it learns on A_2. Therefore, $P_{A_1}(f) \leq P_{A_2}(f)$. $\blacksquare$

It still might be the case that $A_1 \subsetneq A_2$ and $P_{A_1}(f) = P_{A_2}(f)$. For example, let f be defined by $f(x) = 0\ \forall x \in X$. Let $A_1 = \{0, 1, ..., 2^n - 2\}, A_2 = \{0, 1, ..., 2^n - 1\} = X$ where $n > 1$. Then $A_1 \subsetneq A_2$ yet $P_{A_1}(f) = P_{A_2}(f)$ because the only function that agrees with f on A_1 and has complexity at most that of f is f itself.

This points out the uncertainty when dealing with complexity and learning.

For example, it is often helpful to have a "triangle inequality" with a measure (here our measure is complexity — $C(f)$). With this measure, we might hope that for all f and g

in $F, C(f+g) \leq C(f) + C(g)$. However, this is not the case. Note that in this case the $+$ used in adding the functions is the $+$ of Boolean addition where all other sums are as expected except that $1 + 1 = 1$. Consider the following table:

x	$f(x)$	$g(x)$	$(f+g)(x)$
00	0	0	0
01	1	0	1
10	0	1	1
11	1	1	1

Note that $C(f) = C(g) = 0$ since both are projection functions, yet $C(f+g) = 4$. Thus, in this example, $C(f+g) > C(f) + C(g)$.

Perhaps a "reverse triangle inequality" holds. Perhaps it is true that for all f and $g, C(f+g) \geq C(f) + C(g)$. Again, no.

Let f be any function whose cost is not zero. Let $g = \overline{f}$, the complement of f (the function that is zero when f is one and is one when f is zero). Then for all $x, (f+g)(x) = 1$. So that the cost of $f+g$ is zero. Hence in this case we have $0 = C(f+g) < C(f) < C(f) + C(g)$.

Let f_1, f_2, g_1, and g_2 be as defined below.

x	$f_1(x)$	$f_2(x)$	$g_1(x)$	$g_2(x)$
00	0	0	1	1
01	0	0	0	0
10	0	0	1	1
11	0	1	0	1

If we let $d(f,g) = |\{x : f(x) \neq g(x)\}|$, and let $min(f) = min(|\{x : f(x) = 1\}|, |\{x : f(x) = 0\}|)$ (sometimes called "the number of minterms of f"), then $d(f_1, f_2) = d(g_1, g_2) = 1, C(f_1) < C(f_2), C(g_1) < C(g_2), min(f_1) < min(f_2)$, and $min(g_1) > min(g_2)$. So, two functions may be close to one another, have different costs, and we can not resolve which has the lower cost just by counting the number of minterms.

There is not total despair in this area, however, as several bounds relating these ideas can be found in [4].

We now return to the problem of finding the probability that a sample leads to FERD correctly identifying f. Recall that if $A \subseteq f, S \cap C = \{g \in F : g(x) = f(x)\ \forall x \in A$ and $C(g) \leq C(f)\}$. When we speak of more than one function, we will write $(S \cap C)_f$ to avoid confusion. Write $P(A \longrightarrow f)$ to mean the probability that given sample A, FERD learns f.

A projection function is a function of the form $f(x_1, x_2, \ldots, x_i, \ldots, x_n) = x_i$ for all $x \in X$ where i is fixed and $1 \leq i \leq n$. Here we write x in its binary form so that each coordinate is 0 or 1. Thus, a projection function's value at a number is determined by the ith coordinate of the number. When f is a projection function, we take $C(f) = 0$ and $C(\overline{f}) = 2$ by convention.

PROPOSITION 4. *If f is not a projection function then given any $A \subseteq F$, $P(A \longrightarrow f) = P(A \longrightarrow \overline{f})$.*

PROOF: Since f is not a projection function, $C(f) = C(\overline{f})$. So that $o \in (S \cap c)_f \Leftrightarrow \overline{o} \in (S \cap C)_{\overline{f}}$. That is, the structures of $(S \cap C)_f$ and $(S \cap C)_{\overline{f}}$ are dual. Since FERD's learned function is a function in the respective sets with minimal cost, the statement follows. ∎

Now define $C(f, A, i) = \{g \in F : C(g) \leq C(f)$ and $|\{x \in A : g(x) = f(x)\}| = i\}$.

THEOREM 5. *If f is not a projection function and $A \subseteq f$, then for all $k, 0 \leq k \leq a = |A|$, $|C(f, A, k)| = |C(f, A, a-k)|$.*

PROOF: Let k be as stated and let $g \in C(f, A, k)$. Then $|\{x \in A : g(x) = f(x)\}| = k$, so $|\{x \in A : g(x) \neq f(x)\}| = a - k$. But this last set is the same as $\{x \in A : g(x) = \overline{f}(x)\}$, so $|\{x \in A : g(x) = \overline{f}(x)\}| = a - k$. However, if $g = \overline{f}$ on a set of size $a - k$ and $C(g) \leq C(f)$ then $\overline{g} = f$ on a set of size $a - k$ and $C(\overline{g}) \leq C(f)$. Thus, $|C(f, A, k)| = |C(\overline{f}, A, a-k)| =$

$|C(f, A, a-k)|$. ∎

There is one specific result from this Theorem that is worth mentioning. The number of functions that agree with f on a sample and have complexity at most that of f is the same as the number of functions that agree with f nowhere on that sample and have complexity at most that of f, when f is not a projection function. The truth of the previous sentence can be demonstrated by letting $k = a$ in the previous Theorem. We have shown the following.

COROLLARY 6. *Let f and A be as above. Then $|C(f, A, a)| = |C(f, A, 0)|$.*

The Theorem gives us some way of counting the functions that agree with a given function on a fixed sample. We would like to know this number because it lies at the heart of the manner in which FERD operates.

PROPOSITION 7. *If f is not a projection function and $A \subseteq f$, then $|S \cap C| =$*

$$2\sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)| \quad \textit{if a is odd,}$$

$$2\left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k|\right) + |C(f, A, a/2)| \quad \textit{if a is even.}$$

PROOF: Note that $C(f, A, k) \cap C(f, A, j) \neq \emptyset$ if and only if $k = j$. For if g agrees with f on precisely k numbers in A then it can not agree with f on precisely j numbers in A unless $k = j$. Therefore,

$$\begin{aligned} |S \cap C| &= |\{g \in F : g(x) = f(x)\ \forall x \in A \text{ and } C(g) \leq C(f)\}| \\ &= |\bigcup_{0 \leq k \leq a} C(f, A, k)| \\ &= \sum_{k=0}^{a} |C(f, A, k)| \end{aligned}$$

The last equality follows from the argument above. If a is odd, we are summing an even number of terms and

$$\begin{aligned}
|S \cap C| &= \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)| + \sum_{k=\frac{a+1}{2}}^{a} |C(f, A, k)| \\
&= \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)| + \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, a-k)| \\
&= \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)| + \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)| \\
&= 2 \sum_{k=0}^{\frac{a-1}{2}} |C(f, A, k)|.
\end{aligned}$$

The next-to-last statement follows from the previous Theorem. If a is even, we have

$$\begin{aligned}
|S \cap C| &= \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k)| \right) + |C(f, A, a/2)| + \sum_{k=\frac{a+2}{2}}^{a} |C(f, A, k)| \\
&= \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k)| \right) + \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, a-k)| \right) + |C(f, A, a/2)| \\
&= \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k)| \right) + \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k)| \right) + |C(f, A, a/2)| \\
&= 2 \left(\sum_{k=0}^{\frac{a-2}{2}} |C(f, A, k)| \right) + |C(f, A, a/2)|. \quad \blacksquare
\end{aligned}$$

This gives us a start on counting the cardinalities of the sets in which FERD does its learning.

We can generalize the notion of $C(f, A, k)$ a bit, and investigate the following: $\overline{C}(f, A, k) = \{g \in F : C(g) \leq C(f) \text{ and } |\{x \in A : g(x) = f(x)\}| \geq k\}$.

So that, $\overline{C}(f, A, k) = \bigcup_{i=k}^{n} C(f, A, i)$. Now it is not the case that given f and A, $\overline{C}(f, A, k)$ and $\overline{C}(f, A, j)$ are disjoint when $j \neq k$. In fact, from the definition, it can be seen that if $j \leq k$, then $\overline{C}(f, A, j) \supseteq \overline{C}(f, A, k)$.

It can be noted that if $A_1 \subseteq A_2 \subseteq f$ and $0 \leq k \leq a$, then $\overline{C}(f, A_1, k) \subseteq \overline{C}(f, A_2, k)$. For if $g \in \overline{C}(f, A_1, k)$ then $C(g) \leq C(f)$ and $|\{x \in A_1 | g(x) = f(x)\}| \geq k$. Since $A_1 \subseteq A_2$, $|\{x \in A_2 : g(x) = f(x)\}| \geq k$. Therefore, we can say that $\overline{C}$ is what is called homogeneous in A and anti-homogeneous in k. No nice inequalities can be found in the f-coordinate, and no inequalities can be found for the original $C(f, A, k)$ function.

Note that given f and A, the function learned by FERD, g, will be in $C(f, A, a)$ and $\overline{C}(f, A, k)$ for each k.

If FERD identifies f, then $f \in S \cap C$. For the moment, let us assume that FERD has identified f and that $S \cap C = \{f\}$. There are two reasons for doing so: 1) this is the more desirable situation since if $\{f\} \neq S \cap C$, then FERD is only finding f by chance and another learning experiment done with the same function might not produce f, and 2) in the experiments performed with FERD, when FERD correctly identifies f, it appears that $S \cap C = \{f\}$.

Now let $L = \{A \subseteq f : (S \cap C)_f \longrightarrow g \text{ where } C(g) < C(f)\}, E = \{A \subseteq f : (S \cap C)_f \longrightarrow g \text{ where } C(g) = C(f)\}$. Then under our simplifying assumption, $P(S \cap C = \{f\}) = P(\text{sample gives no other function as complex or less complex than } f \text{ in } S \cap C) = P(\text{sample gives no function less complex than } f) - P(\text{ gives no function as complex as } f) = P(\bar{L}) - P(A \cap \bar{L}) = P(\bar{L}) - P(A|\bar{L})P(\bar{L}) = P(\bar{L})(1 - P(A|\bar{L})$. Since $1 - P(A|\bar{L}) \leq 1$, we can deduce that as the complexity of the desired function strictly increases, the probability that a sample will correctly identify f decreases, although perhaps not strictly. Therefore, as might be predicted, FERD has the property that it is harder for it to learn a more complex function than a simpler function. As an example of FERD-learning, we calculate the chance that FERD identifies f when f is a constant function -- that is, either $f(x) = 0$ for all x or $f(x) = 1$ for all x. In this case, if we make the simplifying assumption above, there is really nothing to calculate so we will assume no extra properties of FERD. Also, if we assume that the constant function is the simplest function possible, then FERD will always make the correct identification because there can be no other function that agrees

with f and has complexity no more than f. However, with the cost-complexity defined by Ross in [4], the projection functions have the same cost as the constant functions — a cost of 0.

Let f be a constant function. Then $\forall A \subseteq f, f \in S \cap C$ because regardless of the sample chosen, it is always the case that $f = f$ on that sample and $C(f) \leq C(f)$. If $|S \cap C| = 1$, then $S \cap C = \{f\}$ and FERD will learn f. If $|S \cap C| = 2$, then FERD learns f half the time and learns the other function half the time. Similarly, given that $|S \cap C| = m$, $P(\text{FERD identifies } f) = \frac{1}{m}$. When f is a constant function, the only functions besides f that could possibly be in $S \cap C$ are the projection functions and the other constant function, $\bar{f}$, since they are the only functions whose cost is no larger than the cost of f. However, the other constant function will not be learned by FERD since it does not agree with f on any sample from f. So we can find the probability that a sample results in FERD identifying f by finding the probability that a sample of f leads to k projection functions that agree with f on the sample.

Let k be the number of projection functions that equal f on the sample A. Then, $P(\text{FERD learns } f) = \sum_{k=0}^{n} \frac{1}{k+1} P(|S \cap C| = k+1) = \sum_{k=0}^{n} \frac{1}{k+1} P(\text{exactly } k \text{ projection functions agree with } f \text{ on } A)$.

Assume that $f(x) = 0 \; \forall x \in X$. Write p_i for the ith projection function (the function that maps an x-value to its ith coordinate.) Thus, $p_i = f$ on $A \Leftrightarrow (x)_i = 0 \; \forall x \in A$. So that exactly k projection functions agree with f on A exactly when the coordinates corresponding to those projection functions are 0 for each element of A and for each of the other coordinates there is at least one element of A that is 1 in that coordinate.

Write $x_i = \bar{0}$ on A to denote that the ith coordinate of x is 0 for each x in A. What we have determined then is that $P(\text{FERD learns } f) = \sum_{k=0}^{n} \frac{1}{k+1} P(\text{for exactly } k \text{ coordinates, } x_i = 0 \text{ on } A)$. There are exactly $C(n, k)$ ways to choose the k coordinates (where $C(n, k)$ denotes the combination of n things taken k at a time). Once they are chosen, A must come from the set where each of these coordinates is 0 on the entire set. There are 2^{n-k}

such numbers so that we have $C(2^{n-k}, a)$ ways to choose A if we want at least k coordinates to be $\bar{0}$. (Note that there are $C(2^n, a)$ ways to choose A with no restrictions on A.) We must always assume that $a \leq 2^{n-k}$ otherwise no such A is possible and we take $C(2^{n-k}, a)$ to be zero.

To have exactly k coordinates where each $x_k \in A = \bar{0}$ we must ensure that for each coordinate, i, not equal to one of the k coordinates where $x_k = \bar{0}$, we have at least one x in A where $x_i = 1$. There are $n-k$ coordinates of this type. For convenience in notation, write these coordinates as $i = 1, 2, ..., n-k$. We would like to determine P(for the $n-k$ coordinates, i, $x_i \neq \bar{0}$ on A) $= P(x_1 \neq \bar{0}$ on A and $x_2 \neq \bar{0}$ on A and ... and $x_{n-k} \neq \bar{0}$ on A). These calculations would be fairly simple if the events we are concerned with are independent, but they are not. To determine the desired probability, we use the fact that it is equal to $1 - P(x_1 = \bar{0}$ on A or $x_2 = \bar{0}$ on A or ... or $x_{n-k} = \bar{0}$ on $A) =$

$$1 - \left(\sum_{i=1}^{n-k} C(n-k, i)(-1)^{i+1} P(\text{at least } i \text{ of the } n-k \text{ coordinates } = \bar{0} \text{ on } A) \right).$$

The expression inside the sum sign is obtained using basic probability rules — especially the union rule — which can be found in [2], for example and the fact that $P(x_1 = 0) = P(x_2 = \bar{0}) = ... = P(x_{n-k} = \bar{0})$.

From earlier work, we know that the probability that at least i of the $n-k$ coordinates are $\bar{0}$ on A is $\dfrac{C(2^{n-k-i}, a)}{C(2^{n-k}, a)}$. So that for the $n-k$ coordinates, $i = 1, ..., n-k$, $P(x_i \neq \bar{0}$ on $A) =$

$$1 - \frac{\sum_{i=1}^{n-k} (-1)^{i+1} C(n-k, i) C(2^{n-k-i}, a)}{C(2^{n-k}, a)}.$$

And the probability that FERD correctly identifies f is

$$\frac{\sum_{k=0}^{n} \frac{1}{k+1} C(n, k) \left[C(2^{n-k}, a) - \sum_{i=1}^{n-k} (-1)^{i+1} C(n-k, i) C(2^{n-k-i}, a) \right]}{C(2^n, a)}.$$

We summarize this in the following.

PROPOSITION 8. *If f is a constant function, then the probability that FERD correctly identifies f is*

$$\frac{\sum_{k=0}^{n} \frac{1}{k+1} C(n,k) \left[C(2^{n-k}, a) - \sum_{i=1}^{n-k} (-1)^{i+1} C(n-k, i) C(2^{n-k-i}, a) \right]}{C(2^n, a)}$$

when $a \leq 2^{n-1}$. If $a > 2^{n-1}$, then the probability that FERD correctly identifies f is 1.

We calculate this probability for specific values of n and a. We write $P(F)$ for the probability that FERD identifies f. When $n = 1, f : \{0,1\} \to \{0,1\}$ so $a = 1$ and $P(F) = \frac{1(2-1)+1/2}{2} = 3/4$. When $n = 2$,

$$\begin{array}{ll} a = 1 & P(F) = 7/12 \\ a = 2 & P(F) = 5/6 \\ a = 3, 4 & P(F) = 1 \end{array}.$$

When $n = 3$,

$$\begin{array}{ll} a = 1 & P(F) = 15/32 \\ a = 2 & P(F) = 5/7 \\ a = 3 & P(F) = 25/28 \\ a = 4 & P(F) = 137/140 \\ a > 4 & P(F) = 1 \end{array}.$$

When $n = 4$,

$$\begin{array}{ll} a = 1 & P(F) = 31/80 \\ a = 2 & P(F) = 5/8 \\ a = 3 & P(F) = 109/140 \\ a = 4 & P(F) = 841/910 \\ a = 5 & P(F) = 38/39 \\ a = 6 & P(F) = 143/144 \\ a = 7 & P(F) = 714/715 \\ a = 8 & P(F) = 6434/6435 \\ a > 8 & P(F) = 1 \end{array}.$$

In table form with each (n, a)-entry corresponding to $P(F)$ where $f : \{0, 1, \ldots, 2^n - 1\} \to \{0, 1\}$ is a constant function and $a = |A|$,

	1	2	3	4	5	6	7	8
1	.75	1	--	--	--	--	--	--
2	.583	.833	1	1	--	--	--	--
3	.469	.714	.893	.979	1	1	1	1
4	.388	.625	.779	.924	.974	.993	.999	.999

The table shows patterns that are true in general. For fixed f and fixed sample size (same column), the larger the domain the smaller the chance of learning f. For fixed f and fixed domain size (same row), the larger the sample, the larger the chance of learning f.

During the summer it was found that the ADA function decomposition algorithm (AFD) that is a prime ingredient in FERD was declaring variables/coordinates to be vacuous that really weren't. A variable is called vacuous if its value has no bearing on the function output. More precisely, x_i is a vacuous variable for $f : x_1x_2x_3...x_i...x_n \rightarrow \{0,1\}$ if $f(x_1x_2...x_{i-1}0x_{i+1}...x_n) = f(x_1x_2...x_{i-1}1x_{i+1}...x_n)$ for all values of $x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n$. For example, for the function below, x_1 and x_2 are vacuous variables (where $x = x_1x_2x_3$).

x_1	x_2	x_3	$f(x)$
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

To see if AFD was misidentifying vacuous variables, we counted the number of binary functions that had vacuous variables. This way we know the proportion of functions with vacuous variables to expect in a random sample.

Let $V = \{f \in F : f$ has at least one vacuous variable $\}$. We seek $|V|$. Now, $|V| = |\{f \in F : f$ has at least one vacuous variable $\}| = |\bigcup_{1 \leq i \leq n}\{f \in F : f$ has at least i vacuous variables $\}|$. So once again we use the union rule to count the desired set. If f has at least i vacuous variables then in effect f is a function on $\{0, 1, ..., 2^{n-i} - 1\}$. There are $2^{2^{n-i}}$ such functions. Therefore,

$$|V| = \sum_{i=1}^{n} C(n,i)(-1)^{i+1}2^{2^{n-i}}.$$

This number is calculated for some values of n below.

n	$\lvert V \rvert$
1	2
2	6
3	38
4	942

The table shows that $|V|$ does grow quickly but so does $|F|$. In fact, we expect $\frac{|V|}{|F|}$ to be on the order of $\frac{n}{2^{2^{n-1}}}$, a very small proportion.

RECOMMENDATIONS

Machine-learning is still a very new and largely unexplored area. Often an algorithm is presented that allows machine-learning for a specific type of function, for example, letter recognition. The FERD algorithm is different in that it attempts to learn any function regardless of its type. This strength makes it a powerful tool whose workings are still to be found. Progress has been made but must still be made.

We would like to try FERD on some larger functions — ones with larger domains — but currently that is impractical.

A study has been done on how decomposed function cardinality relates to Kolmogorov complexity — a popular measure of complexity [3].

Recently, the concept of Vapnik-Chervonerkis dimension has come to our attention, especially through [1]. Its title suggests why it is of interest because it is desirable to know how large a sample we need to take in order to guarantee efficient learning.

We would like to relate our work to existing work to 1) indicate to outsiders the framework within which FERD is working and 2) to perhaps use some general theory that may apply to our algorithm. Presently, we have yet to find that link. Either the theory is too specific and so does not apply to FERD or it is too general to be of use.

However, make no mistake about the fact that FERD has outperformed more conventional and more well-known machine-learning systems. We will continue both to try to find out why and to broaden its application.

REFERENCES

1. Ehrenfeucht, Andrzej et al. *A General Lower Bound on the Number of Examples Needed for Learning*. Information and Computation **82** (1989), 247–261.

2. Hogg, Robert V. and Eliot A. Tanis, "Probability and Statistical Inference 3d ed.," Macmillan Publishing Company, New York, 1988, p. 57.

3. Li, Ming and Paul Vitanyi, *Two Decades of Applied Kolmogorov Complexity*, Proceedings: Structure in Complexity Theory (1988), 80–101, IEEE.

4. Ross, Tim , Mike Noviskey, Tim Taylor, and Dave Gadd. *An Engineering Paradigm for Algorithm Design*. Technical Report WL–TR–91–1060 (1991).

A Formal Process Model for Software Re-engineering: The Analysis Phase

Eric J. Byrne and David A. Gustafson PhD

Report TR-CS-91-12

Department of Computing & Information Sciences
Nichols Hall
Kansas State University
Manhattan, KS 66506

This research was supported by the U.S. Air Force - Universal Energy Systems RIP program. Contract # F19620-88-C-0053/SB5881-0378

Abstract

This report presents the results of research into the software re-engineering process. The demand for successful software re-engineering techniques is growing. Research into software engineering tends to focus on technical aspects of the problem. However, effective technical solutions must fit into an overall framework that represents the entire software re-engineering *process*. While technical solutions are needed, there is more to the process of re-engineering software than just technical work.

The software re-engineering process can be divided into numerous tasks. These tasks fall into three groups based on the type of work done in each task. The three groups are: management, technical, and support. Much of the research into software re-engineering focuses on technical aspects of re-engineering. While such research is important, the key to a successful re-engineering project is planning and management.

The goal of this research is to model the software re-engineering process. This report presents a new model of the re-engineering process that adds several more phases to the traditional re-engineering model. This report examines the Analysis and Planning phase of the re-engineering process and presents a formal specification of the information objects that are manipulated during this phase. The tasks that occur during this phase are also discussed.

Chapter 1

Research Overview

There is a growing interest and need for successful software re-engineering methodologies and techniques. Research into software re-engineering tends to focus on technical solutions. While sound technical procedures for re-engineering are needed this emphasis ignores other aspects of the software re-engineering problem. One lesson learned from experience in software development is that sound management and planning are necessary for a project to succeed. This same principle is true for any software re-engineering project. This report examines the *process* of software re-engineering. The planning phase of the re-engineering process is examined in detailed.

1.1 Introduction to Research

The purpose of the current research is to study the **process** of software re-engineering. The motivation for this research is driven by a simple question: How can a software re-engineering *project* be accomplished? If a re-engineering project was studied and all the work done within the project was noted and recorded, what tasks would be found? What information would be created? Manipulated? How would the tasks be ordered? How would the project be organized? What skills would be needed? What tools would be useful? What problems would be encountered? How would they be solved? By asking one question, many questions appear.

To answer these questions, this research seeks to produce a formal process model that specifies the overall re-engineering process and the subprocess activities such as software reverse engineering and re-design. The questions driving this objective are: ***What information must be produced?*** and ***When can it be produced?*** The first question focuses attention on the information used within the re-engineering process. The properties and characteristics of this information must be stated. Relationships between information groups must be clarified and expressed. The second question focuses attention on the tasks that create and/or manipulate this information. Properties of these tasks must be stated and the relationships between tasks must be expressed. It is also necessary to declare the assumptions on which the process model is based, and the re-engineering capabilities that are modeled.

What is a software process? A software process is the collection of related activities, seen as a coherent process, used to accomplish some task such as software development, testing, or

maintenance [Dowson86]. What is a software process model? A software process model represents attributes for a particular software process and is specific enough to allow reasoning about the process.

What is the significance of a process model? Software process models are seen as important vehicles for understanding, evaluating, reasoning, and improving software processes [ISPW88]. Processes are hard to understand and reason about, while process models, as static objects, are far easier to comprehend. A process is a vehicle for doing a task, while a process model is a specification of how the task is to be done. A software process model represents key relationships among a variety of types of software objects. Typical objects that are modeled are activities, data objects, tools, and user roles within the process. Thus, a formal process model can be used to provide an understanding of the essential properties of the re-engineering task.

The research reported here is a continuation of research done by Byrne [Byrne90]. In his original report, a descriptive process model of the technical aspect of software re-engineering was presented. The major technical tasks covered were:

Design recovery The reconstruction of a system design from source code.

Re-design The alteration of the reconstructed design to satisfy new criteria.

Re-implementation The implementation of the altered design.

Re-documentation The updating of the system documentation to reflect differences between the target system and the existing system.

Strangely, in the first report a fifth technical sub-process, testing, was omitted. The descriptive process model focused on the tasks within the re-engineering process. The inputs and outputs for each task were identified. The ordering of tasks within a hierarchy was presented. The flow of information between tasks was also given.

During the research reported here it was realized that the full software re-engineering process has three major aspects or areas of concern. These areas are project management, technical work, and project support. The first report considered only the technical aspect of the re-engineering process. By broadening the scope of the re-engineering process it is possible to obtain a more complete and accurate model of the process. What quickly becomes almost overwhelming is the variety and volume of information that is generated during a project. There are many information objects and operations on these objects. The relationships between objects is complex. By defining a model of the process it is possible to clarify the properties of these information objects and their interrelationships.

The realization that only one aspect of the re-engineering process had been covered in the first report led to a dilemma. The original intent of this stage of the research was to improve and build on the initial process model presented in the first report. However, doing so would mean that the other two aspects would continue to be ignored. It was decided that since the goal of this research is to model the entire re-engineering process, more of the process should be explored. Therefore, the research reported here examines the overall software re-engineering process. Once the overall structure of the process was expressed, the work focused on one phase of the process. The analysis

and planning phase, which is the first phase of a re-engineering project, was examined in detail. The results of that examination are presented in this report.

The second objective of this research is to identify a suitable notation to express the developed process model. To represent a formal process model that can be reasoned about, a process modeling language is used. Such a language has a formally defined syntax and semantics that permit models expressed in the language to be analyzed and properties of the model to be discussed.

In the first report each task was carefully explained in English using a template to format the information. The flow of information between tasks was modeled using dataflow diagrams. While this conveyed information about the ordering of tasks, it gave little information about the characteristics and properties of the information being manipulated.

For this report several process modeling languages were examined. Eventually, it was decided that the level of representation offered by these languages was too detailed. These process programming languages were aimed at implementing a process model. The current research is focused on the re-engineering process at a higher level of abstraction. Finally, it was decided to use the formal specification language Z [Spivey89]. The Z language was designed to state the properties of programs and data, while avoiding design or implementation details. While not developed for process modeling, it has proven to be usable in specifying the properties of information objects and tasks within a process model.

1.2 Software Re-engineering

Most software organizations report a growing backlog of maintenance work. In many organizations resources are being diverted from new development projects to support the maintenance of existing systems. It is estimated that many software organizations spend between 60% - 75% of their effort on maintenance. Much of this is work on outdated code that is difficult to maintain, but is considered valuable by its users.

As a software system is maintained over a period of years the logic and structure of the system deteriorates. Eventually, the system becomes difficult to maintain and requires *guru* personnel who have accumulated a deep familiarity with the system and its quirks. At this point re-engineering of the software is often considered.

Software re-engineering, also called software re-juvenilization, reclamation, or recycling, is gaining attention as the base of installed software ages and becomes in need of replacement. Most organizations have invested millions of dollars in their software and are not willing to simply throw it out and build completely new systems. These organizations have a great deal of experience invested in their systems and are more willing to modernize a system than to obtain a new one. Software re-engineering addresses this problem.

Software re-engineering is necessary when a software system reaches the point where it must be retired or rewritten. Re-engineering is selected when the system is seen as valuable but too difficult to maintain in its current state. This is a problem faced by software systems that originated 10, 20, or more years ago. Old age is not necessarily the only reason to re-engineer a system. Here are some typical reasons:

- An organization has purchased a new computer. The new environment does not support the programming language in which parts of the existing system are written, or the system must be rewritten to use features provided by the new environment.

- A software system that has been around for many years has become too expensive to maintain. Such systems have generally evolved far past their original functionality and the code has become difficult to understand or change.

- An organization has decided to start using modern software engineering practices, but unfortunately, most of the organization's work deals with already existing code that does not have the benefit of these new practices.

- An organization plans to purchase modern CASE tools that support development and maintenance. Unfortunately, such tools tend to work only on code initially developed with the tool.

- An organization has decided to switch to a modern programming language, and wants to translate their existing code into the new language. For example, the DOD mandate on ADA.

Software re-engineering begins with an existing software system and produces a new modernized version of that system. Chikofsky and Cross defined software re-engineering as [Chikofsky90] :

> The examination and alteration of a subject system to reconstitute it in a new form and the subsequent implementation of the new form.

The term "re-engineering" implies some type of transformation. A software re-engineering process should start with an existing system and generate a new high-quality system with valid documentation. The definition of such a process can begin by comparing the characteristics of existing systems (input to the process) with target systems (output from the process). In this way, functional requirements can be identified. In addition, several process goals and constraints are also apparent.

The software re-engineering process must provide the capability to produce a new system that is written in a different programming language than the existing system. One reason for re-engineering a system can be to switch to a modern programming language. Perhaps parts of the system are written in different languages and the new version is to be written in only one language.

The process must provide the capability to re-design the system architecture. An existing system typically was designed before modern design techniques came into practice. Modern techniques such as information hiding and object-oriented design are intended to produce systems whose architecture reflects a model of the application domain. Designs produced by these techniques tend to be easier to comprehend, maintain, and change.

The process must provide the capability to incorporate new requirements into the system. This means changes to the functionality of the system. This may refer to the ability to work on a different machine architecture, interface with new equipment or systems, changes to the user interface, addition of new capabilities, etc. It can also mean the removal of portions of the system

that support requirements that have been deleted. Or it can mean changes to the system to satisfy changes to existing requirements.

The process must provide the capability to generate structured source code even when the existing code is unstructured. This refers to the use of control-flow constructs used in the body of a routine.

The process must provide the capability to restructure data. The capability to restructure data aids in improving the understandability and maintainability of the code. Data restructuring includes changing data structure definitions, removing duplicate data items, clarifying type information, organizing data handling procedures and re-organizing data files (data re-engineering).

The process must provide the capability to document the system on both the program and design level. Program documentation includes information such as where an item is defined and where it is used, who-calls, who-is-called, item descriptions, etc. Design documentation represents the system at a more abstract level, and shows the system architecture and processing characteristics. Design documentation should document both high and low level design.

The process must preserve the correctness of the existing system. Preserving the correctness means that the functionality of the system has been preserved, with the exceptions of changes to the system specified by requirement changes. The requirement that correctness be preserved requires that the process either guarantees preservation or provides a method to verify it.

The process must provide the capability to replace existing program item names with meaningful names. This is the ability to rename any variable, procedure, function, macro, data type, etc. In addition to replacing a name, the process must aid in locating and replacing all occurrences of that name. This includes both source code and documentation.

The process must support the generation of a system that is organized and written to support understandability. This is a goal and not a functional requirement. Many of the requirements above, such as use of modern design techniques, control-flow structuring, and use of meaningful names will help achieve the goal of understandability.

The process must support the generation of source code that conforms to an in-house standard. Such a standard places constraints on the presentation style of the source code, naming conventions, restrictions on implementation language constructs, etc. The process should help verify that such constraints are not violated.

Examining these requirements it becomes clear that the re-engineering process must be able to manipulate a representation of the system. Further, this representation must be more abstract than a source code representation. This can be seen by examining the requirement that the process must be able to manipulate the system architecture or high-level design. Manipulating the system architecture by changing source code is difficult, partly because such manipulations must consider the implementation details of the system. Modifications can be more easily made to a design representation of the system. A design representation for an existing system can be constructed using reverse engineering techniques.

How should software re-engineering be done? The U.S. Federal Conversion Support Center, Office of Software Development issued a set of guidelines for establishing software improvement programs (SIPs) ([OSD/FCSC83], [Houtz83]). These guidelines establish "what needs to be done," [illegible]

modernization of existing software to maximize its value, quality, efficiency, and effectiveness. Modernization includes those activities necessary to upgrade the software and its engineering techniques to current or state-of-the-art levels. A SIP preserves the value of past software investments, and at the same time, increases the reliability, efficiency, portability, and maintainability of the software to create an improved foundation on which to maintain older systems.

The SIP approach is intended to establish a program for improving software within an organization. A SIP is not a one-shot effort, it establishes an evolutionary approach to software improvement. The software improvement approach may be used to improve existing software or build new systems from existing software. When working with existing systems there are several factors that influence or constrain the work. The characteristics of the existing system must be considered. There may be other operational systems that are integrated with the existing system. The testing integrity of the current system may need to be preserved while moving to a new or improved system. An evolutionary approach may be necessary because of a need for usable implementations at each stage of the improvement process.

Another approach for improving software is Structured Retrofit which is based on restructuring techniques[Lyons81]. In one sense, restructuring is a limited form of re-engineering. Typical restructuring techniques focus on control-flow restructuring of source code. This is the transformation of unstructured source code into structured code by removing GOTO's and other unstructured control-flow constructs.

The basic philosophies behind both structured retrofit and re-engineering are similar. The act of changing software tends to destroy or obscure its structure. This makes software progressively more resistant to change. By examining the processes embodied in these and other approaches it is possible to discover tasks and information that software re-engineering must also incorporate.

1.3 Guide to Report

The focus of many research projects in software re-engineering is the development of techniques and tools that support a programmer. What is missing is an investigation of the deeper properties of software re-engineering. The development of supporting tools without an understanding of the real problems that need to be addressed will not lead to satisfactory solutions that can reduce the complexity of a task.

What is needed is a deeper understanding of the characteristics of software re-engineering. This deeper insight can be provided by two lines of investigation. The first is to examine and model the re-engineering process. The second is to establish a formal basis for re-engineering techniques. This research focuses on the first line of investigation: examine and model the re-engineering process.

The first objective of the research reported here is to identify a suitable language for representing the software re-engineering process model. The second chapter presents the concept of software process modeling and process modeling languages. The language chosen to express the process model is Z. At the end of chapter 2 a brief description of Z is given. Appendix A also summarizes the Z notation used in this report.

The second objective of the research is to model the software re-engineering process. Chapter 3 presents the software re-engineering process and its major phases. This is a high-level discussion

of the activities involved with conducting a re-engineering project.

Within the software re-engineering process the first major phase, project analysis and planning, was examined in detail. Chapter 4 contains a formal specification of the information and operations used within this phase. The tasks that must be done to successfully plan and prepare a re-engineering project are also discussed.

Chapter 2

Process Modeling and Languages

The goal of this research is to develop a model of the software re-engineering process. A model that contains enough detail to guide humans through the process, providing information about major and minor tasks, their sequencing, and their steps. The process model developed must accurately describe the process, otherwise it will be of little value.

An issue that must be resolved is how to represent the software re-engineering process model. English descriptions are neither precise or concise. Diagrams omit too much important information. A notation or language for expressing the model is required. There is a subfield of software engineering dedicated to researching the software process. The specification and development of process modeling languages is an area of research within this field.

2.1 Software Process Research

To study software processes it is necessary to realize that they exist. Programmers use a process to develop software. They follow a process for specifying requirements, a process for designing software, a process for implementing it, for testing, for fixing errors, for porting software, etc. The result of a software process is a software product. Organizations often have procedures or guidelines that are followed when developing or maintaining software. Osterweil said, "...humans improvise processes on the fly, maintain them mentally, modify them as necessary and guide others in their effective use..." [Osterweil87]. By studying software processes their existence is recognized. These processes must be made visible. They must be described formally, discussed, reasoned about, and improved.

A software process model is a means of expressing a software process. Researchers who study process modeling agree that if you are concerned about the quality of software products, then you need to be concerned with the process by which they are produced [ISPW88]. Software process models are seen as important vehicles for understanding and reasoning about software activities [Williams88]. The ability to represent and model a software process is necessary to understand, evaluate, reason, and improve the process. Software process models are necessary to help make software activities more reliable and productive.

Until recently, process models have focused mainly on management aspects [Curtis87]. Boehm

states that software process models are important because they provide guidance on the order of major phases in a project. He believes that the primary functions of a software process model are to determine the order of phases involved in software development and evolution and to establish the transition criteria for progressing from one phase to the next [Boehm88]. This point of view is reflected in software life-cycle models.

The Waterfall model is perhaps the most well known software life-cycle model. The Waterfall model and its variations describe the software process as a series of phases. Each phase ends with the production of an artifact, such as a design document, or source code. The Spiral model [Boehm88] replaces the Waterfall approach with a spiral model of development that uses a risk-driven approach rather than a document-driven or code-driven process. Iivari builds on this to create the Hierarchal Spiral Model that combines the spiral model with the PIOCO model [Iivari87].

The problem with these life-cycle models as software process models is that any model built only on the end points of major phases offers little insight into the actions and events that take place to produce the end point products [Curtis87]. These models enable management to set milestones and measure progress, but provide little information about how software is actually created or modified. These life-cycle models are actually high-level process models that are not elaborated to lower levels of detail. Process models that describe lower levels of detail provide the means to discuss exact procedures.

What is a process? Osterweil defines a process as a systematic approach to the creation of a product or the accomplishment of some task [Osterweil87]. Dowson defines a software process as the collection of related activities, seen as a coherent process subject to reasoning, involved in the production of a software system [Dowson86].

What is a software process model? Dowson defines a software process model as a purely descriptive representation of the software process [Dowson86]. A software process model must represent attributes of a range of particular software processes and be specific enough to allow reasoning about them. Mi gives a different definition. Mi defines a software process model as a prescriptive representation of software development activities in terms of their order of execution and resource management [Mi90]. This difference in definition, are process models descriptive or prescriptive, is currently being debated, but most researchers seem to agree that software process models should be prescriptive.

Processes are hard to comprehend and reason about, while process models, as static objects, are far easier to comprehend. A process is a vehicle for doing a task, while a process model is a specification of how the task is to be done. If a software process model is prescriptive, it can be written as a process program that guides a human through the encoded process. This is the viewpoint held by many and best stated by Osterweil [Osterweil87].

> We suggest that it is important to create software process descriptions to guide our key software processes, that these descriptions be made as rigorous as possible and that the processes then become guides for effective application of computing power in support of the execution of processes instantiated from these descriptions.

Lehman argues against this approach [Lehman87]. He claims that the pursuit of prescriptive models is useless, even meaningless, unless we have comprehensive knowledge about the process.

process so far.

In either case, it is agreed that languages for representing software process models are needed. The questions is whether we need software process modeling languages or software process programming languages.

A software process meta-model is a representation formalism that provides the necessary components to create various types of software process models [Mi90]. Katayama points out that a key software process issue is the choice of formalisms. A formalism must provide a clear and understandable description. It must be capable of evolving over a period of time. It must provide the necessary ability to describe the behavior of the process. Two important behavioral aspects of software processes are concurrent processes and iteration [Katayama89]. Formalisms are needed to represent information about the software product, operations on the software product, sequencing of operations, information about resources including project personnel, equipment, and facilities [Sutton89]. Humans are essential components in the software process and their roles and limitations must also be subject to definition and design [Gamalel-Din88].

The software process is shaped by many factors. The individual programmer is concerned with the process used to accomplish his work. A team of programmers are concerned with how their work interacts. Management is concerned with the sequence of activities that must be done, allocating the necessary resources, and tracking the advancement of the work. Even policies of the organization can effect the software process. Because of these influences a software process model can be viewed from a variety of perspectives [ISPW88]. These viewpoints include life-cycle, organizational, and software process improvement-oriented. The value of any view is to provide insight into the important characteristics of software processes and the issues that must be addressed in modeling these processes. A software process modeling language must be able to support these differing views.

The language used to represent a prescriptive software process model should be enactable. The term "enactable" was adopted by the 4th International Software Process Workshop [ISPW88] to convey the concept of a "running" process and avoiding the connotations of machine execution that terms like "executing" or "interpreting" suggest. A software process model written in an enactable language is called a software process program. The activity of expressing software process models with the aid of programming techniques is called process programming [Osterweil87]. The result of running a process program is typically a software product program. The term product program is used to denote a standard executable program written in a normal programming language such as C, Fortran, or Lisp. A process program must be written in a process programming language.

How does a process programming language differ from a product programming language? In a standard programming language primitive expressions are formed from a set of basic operators that are part of the language. In a process programming language the basic operators are tools and these tools are not part of the language. A tool can be an automatic tool that is invoked under certain conditions or an interactive tool, such as an editor, that allows a human to do some necessary task. Another difference is the data types supported by the language. Most programming languages support a fixed set of primitive types such as character, integer, real, arrays, etc. A process programming language must provide a rich typing system to allow programmers to specify complex types for objects such as design documents, test plans, source code, cross reference information, error reports, change requests, etc.

A prescriptive software process model provides a basis for structuring software environments. Most software environment currently supply a toolbox. Providing a collection of tools from which a user selects and uses what is needed. This is a loosely coupled architecture and provides little means of insuring that tools are used correctly or even at all [Ramanathan88]. There is a growing opinion that an unstructured "bag of tools" does not qualify as a software environment. Dowson defines a software environment as [Dowson86]:

> A coordinated collection of software tools organized to support some approach to software development or conform to some software process model.

The collection of tools is coordinated by an explicit process model. The model is not built into the environment but is represented explicitly in a process programming language and can be changed by the users. The ability to change the process model used by an environment provides much flexibility. The local organization or user can tailor the process model to conform to local work patterns or methods.

2.2 Process Modeling

Why should we develop process models? Software process models provide the means to make visible the processes by which software is developed and maintained. Software process models enable increased understanding of a process and help to highlight important features [Sutton89]. Kellner has identified four objectives for the development of software process models [Kellner89].

1. Enable effective communications regarding the process being modeled.
2. Facilitate reuse of the process.
3. Support evolution of the process.
4. Facilitate management of the process.

Once a process model is developed, how should it be used? This question invokes the debate on whether models should be descriptive or prescriptive. Descriptive models are defined to gain a better understanding of a complex process by exposing some important characteristics, but without actually providing any details on how the process is carried out. Prescriptive models provide more or less firm guidelines on how the different actors of the process will work to achieve the goals. A prescriptive model is, at least implicitly, based on a descriptive one.

A prescriptive software process model can be automatically analyzed and enacted by computers [Osterweil87], such a model is called a process program. This suggests that process programs can serve as a means of specifying how the actions of software tools might be integrated with human activities to support software processes. Thus a prescriptive software process model can be written to define possible (allowable) patterns of behavior between humans and tools in a process [ISPW88]. Enactability means that human beings involved in the software process can receive computer guidance and help in what is an extremely complex activity.

2.2.1 What are we trying to model?

If a prescriptive process model for software design is developed, will it be able to design software when enacted? The answer is no. Ramanathan has pointed out that a large portion of software development is creative and cannot be automated. However, interspersed with this creative activity is much planning, control, status tracking, reporting, and other management activities for which well defined standards exist, but that are difficult to enforce manually [Ramanathan88]. A process program can be used to guide programmers through the process while enforcing management policies, yet providing humans with the necessary tools to accomplish their work.

This represents one possible view of the use of a software process model: the management view. Mi claims that effective process models should address organizational and technical dimensions including:

1. Detailed descriptions of software processes,
2. Their interactions,
3. Management and exception handling during the performance of the software process,
4. Product-specific, organization-specific, and project-specific processes.

Software process models will be used by different groups of people, each with their own view of what information is necessary. Three possible views are technical, managerial, and organizational. Curtis in his study of software develop projects found the following hierarchy of process factors [Curtis87]:

1. Individual concerns
2. Team concerns
3. Project concerns
4. Company concerns

The individual programmer is concerned with how to accomplish his task. What tools should be used? What steps are necessary? How should each step be done? The team is concerned with interactions among the team members. Interaction includes dividing tasks, communication between team members, and possible resource conflicts between members. Project level concerns deal with scheduling project phases, meeting deadlines, allocating necessary personnel and resources to accomplish the project. Finally, the company affects the process by setting company goals, directions, and policies.

There seem to be many possible foci for process models. Curtis believes that the most valuable are those that capture the processes that control the most variance in software productivity and quality. If a process model does not represent the processes that control the largest share of variability within a software process, it can not be helpful in improving productivity and quality [Curtis87].

To accommodate this variety of views is has been suggested that a single representation capturing all the relations among objects and processes be created. Then tools could be used to extract specific views on request [Taylor88]. However, it may be difficult to provide a single description that is rich enough to support all the various views [ISPW88]. Despite these complications, there is general agreement that a software process model must describe the context in which problem-solving decisions are made.

A software process model represents key relationships among a variety of types of software objects. These relationships are dependent on the "view" being used to examine the process model. Software objects are complex. Examples of software objects are requirement specifications, design documents, user manuals, cross-reference information, module interface specifications, test plans, test results, change requests, etc.

A prescriptive process model performs activities on objects. Activities are done with tools. Tools can be fully automated and invoked by the process model, or interactive and used by a human to accomplish a task. A process program shows how the various software tools and objects are coordinated to support a process. Software tools are operators used by a model. They transform software objects. Humans are assigned well defined roles in creating and transforming objects too. The specification of what they do, when they do it, and how they coordinate with others humans and with their tools is embodied in a process program.

There are several dangers in process programming. For example, the process model may not map accurately to real process behavior. A process program must achieve an accurate model of behavior. Existing models of the software development process do not provide enough insight into real development processes to guide research on the most effective development technologies [Curtis87]. One solution is to maintain a history of process usage to form a trace. Analysis of a process trace can pinpoint problems with the process and high-light successful processes used by programmers.

2.2.2 Developing Software Process Models

How is a process model created? Taylor reported that in an effort to gain experience with process programming, process programs were written for various software processes. While the intent was to gain experience with process programming, it soon became clear that it was necessary to develop process requirement specifications and designs first [Taylor88]. He also "discovered" that process designs proved to be more useful in understanding the processes than the process code itself. This agrees with statements by others that a life-cycle for software process programs (or models) exist. It has been proposed that methodologies used for system analysis and design should be considered as suitable approaches to process model development [Kellner89].

2.3 Process Programming Languages

A language for representing process models is necessary. For prescriptive process models the chosen language must be enactable, i.e. it must have an execution mechanism. Unfortunately, the specification of a language depends on an accumulation of experience with process modeling or process

programming. What features does such a language need? Rapid progress in process programming can not be achieved until a suitable language for process programming exists [Osterweil87].

The 4th International Software Process Workshop focused attention on languages and notations in which formal models of software processes could be represented and enacted [ISPW88]. There was a general consensus that process programs are likely to be inherently complex. The concern for expressive range in a process programming language must be balanced against concern that notational complexity does not add to the already high inherent complexity of process models.

Boehm applied process programming to his Spiral Model [Boehm88] and reported the following lessons learned about process programming [Boehm89].

1. A model for the meaning of the process program is essential.
2. Data representation issues are significant.
3. Relationships among data objects are a central issue. Relationships must be made explicit.
4. Data visibility issues are significant.
5. Change of the process during its elaboration is critical (dynamic properties).

Can experience with developing product programming languages be applied to developing process programming languages? Indeed, can product programming languages be used for process programming languages? The general consensus is that while much can be taken directly from product programming languages, these languages are not well suited for process programming.

Osterweil believes that the primary difference between process programs and conventional product programs is that process programs represent programming in a non-traditional application domain [Osterweil87]. Curtis believes that a difference between process programs and product programs is the variability of the underlying process being specified [Curtis87]. In a conventional programming language the basic operators are well defined. In a process programming language the basic operators will be tasks that are process specific and carried out by automated tools or humans. The 4th International Software Process Workshop suggested that the relative differences between process programs and product software include much less code in process programs, and increased amount of instantiation, and more concern for failures in instantiations and composition.

2.3.1 Data Support

Software process programs specify how software objects are to be systematically developed or modified. Software objects include much more than just source code. Other possible objects are requirement specifications, designs, documentation, user manuals, test cases and test results, test plans, object code, executable code, change requests, versions of objects, etc. Software objects are complex data aggregates.

Software processes manipulate information. This information differs from that needed for product programs in several ways such as content, structure, and the properties of the supporting operations. Software information objects are stored in an information base. The list of information and support provided by the information base includes: objects of varying sizes, varying

degrees of persistence, nested transactions, very long transactions, complex and programmable relations among objects, triggering mechanisms, automatic inferencing mechanisms, dynamic types and schemas, multi-user sharing with associated locking mechanisms, versioning, query languages, and partioning and view mechanisms.

Normally, such information is maintained in a database. The connection between a process program and the database that serves as a repository for the objects and relationships included in the software process is important. For a process program this database serves as part of the program store. Because of this the connection between the language and the information base must be considered. The requirements for the information store must be carefully considered. They are richer than for any other known application area.

2.3.2 Survey of Process Programming Languages

Because of the lack of an accepted process programming language, research projects that investigate software processes often develop their own notation or language to express a process model or program. These languages vary greatly in level of expression ranging from high level specification style languages to lower level traditional programming style languages. Languages run the range of styles such as imperative, object-oriented, and rule-based.

The following subsections summarize several process modeling and process programming languages that have been reported in the literature. Available information for these languages range from short articles to fully implemented systems. Many of these languages are experimental. The suitability of each language was considered for use in this research work. The criteria for choosing a language is given in the next section.

IT : The Meta-Model

Wileden proposed a high-level modeling representation called IT [Wileden86]. This meta-model formalism is based on two abstract spaces called the I - information space, and the T - transformation space. Transformations are functions that take information and produce information. For example, a simple Waterfall model can be expressed by dividing the information space into a requirements space, design space, and code space. Transformations can then be defined to take information from the requirements space and generate information in the design space, and similarly from the design space to the code space.

HFSP : Hierarchical and Functional Software Processes

Katayama introduced a formalism for hierarchical and functional software process descriptions called HFSP ([Katayama89],[Katayama81]). HFSP defines software processes through hierarchical activity decomposition. The basic principle of HFSP is to focus on the product-base characterization of activities and their hierarchical decomposition. HFSP is founded on two fundamental concepts, (1) design activity and (2) activity decomposition. Its formalism is derived from attribute grammars. HFSP can be characterized as a specification style language. Process models expressed in HFSP can be enacted.

In HFSP an activity is the unit of task in a software process. An activity is assumed to be completely characterized by its input and output attributes. An activity may be simple or a complex task that must be decomposed into subtasks. If an activity is simple enough to be performed by invoking tools, its execution is left to the human activity of doing the job by using the tools. If it is not the activity must be decomposed into sub-activities. Activity decomposition must specify how an activity is decomposed into other activities and what relationships hold among attributes of the activities involved.

Concurrency is essential for describing design processes for large software products, because software design is usually performed by a group of designers working together. In HFSP, concurrency is expressed through attribute dependency. Nondeterminacy is useful in describing design alternatives. HFSP expresses nondeterminacy through decomposition conditions.

SPM : Software Process Model

Williams presented an approach to software process modeling based on behavioral descriptions of software activities [Williams88]. Behavioral descriptions make it possible to describe a process at any desired level of abstraction. The use of abstraction assists in handling aspects of a process that are poorly understood. SPM style specifications are used for descriptive process modeling.

The Behavioral Model describes a software process as a collection of activities or processes that can take place concurrently and asynchronously. The model does not assume that information is "shared" among activities or the individuals performing them. A process is described in terms of abstractions. The approach is behavioral because descriptions of activities focus on the effects that the activities produce rather than the specific procedures used to produce those effects.

A process model developed using this approach is described by a set of activities. Each activity has a set of preconditions, an action, a set of postconditions, and a set of messages. Activities may be performed by different individuals and at different locations. Activities may be built-up from simple activities. Activities can be ordered sequentially or in parallel.

All preconditions must be satisfied before an activity can be started. At least one postcondition will be true when the activity ends. Preconditions and postconditions are not used to prove particular properties of the software process. Instead, they represent assertions about the properties of various activities and the interconnections among those activities.

An action is a software task that may be performed by invoking a particular tool or sequence of tools or may be carried out by a human without automated assistance. An action may have several preconditions, all of which must be satisfied before that action can be performed. An action may have several postconditions, only one of which will be satisfied when the task is completed.

Messages are used for communication and synchronization between activities. They also provide a visibility mechanism for project management in that messages can be associated with project milestones.

Specification of the particular activities to be included and the order in which they are to be performed defines a particular software process. The set of possible activities forms an "alphabet" for software process description and describes a class of software creation and evolution activities. A particular process is defined using a language (the process description language) over this alphabet.

A given software project then corresponds to an individual string in this language.

A software process description may be written using the constrained expression formalism. A constrained expression describing the possible behavior of a software process consists of a process expression and a set of constraints. The process expression typically describes activities of parts of a process without regard to their potential interactions with other parts. The constraint set imposes restrictions on how events from various parts of the process expression may be combined into possible behaviors. The use of the constrained expression formalism provides several advantages. First, this notation allows a concise description of a software process, including concurrent activities. Secondly, the constrained expression can be analyzed to determine whether a given process description produces a correct, or acceptable sequence of activities.

This approach provides the capability to explicitly model human actions as well as tool invocations. Adding messages makes communication among activities explicit. The formal approach used in SPM makes it possible to analyze software process descriptions using techniques developed for event-based descriptions of software systems.

CML : Conceptual Modeling Language

The TRIAD project uses the CML language (Conceptual Modeling Language) to specify process programs ([Ramanathan88], [Ashok89]). A project model for a TRIAD environment is composed of four related models: a process model, data model, tool model, and user model. CML has features for process, data, tool, and user modeling. The representational requirements of these four models are different from each other. However, there is a substantial amount of interaction between the models. Therefore, all the modeling primitives are unified under one language formalism. This research project is developing a compiler that translates project specific descriptions written in CML into intermediate representations that can be interpreted by generic interpreters that constitute the runtime system of an environment.

The process modeling formalism in CML is intended to represent those well-defined aspects of the software process. The process modeling formalism provides primitives for representing the following features of the software process:

1. The representation of each project related activity as a progression from states to states.
2. The representation of activities at various levels of abstraction.
3. The concept of ownership (responsibility) for an activity.
4. Communication between agents performing activities.

A project specific process model contains definitions of activities and the relationships between them. Each activity has a precondition, a postcondition, zero or more attributes (which include references to database objects and tools that are used during that activity), zero or more subactivities, several local state variables, and an action part containing imperative code. Activities are organized into a hierarchy by subactivity relationships, representing an elaboration of activities to the desired level of detail.

An activity may begin if its precondition is true. The action part of an activity can contain statements to manipulate the database, interact with the user, invoke tools, or manipulate the process model. Some examples of statements that manipulate the process model are create_task, delete_task, wait_user, send_user, start_subactivity, and wait_subactivity.

A formalism for data modeling is built into CML. This formalism is based on an object-oriented semantic data model and has been developed for modeling software objects. Besides providing primitives for modeling the structure of software objects and interobject relationships, a rule formalism is also provided to express domain-specific semantics of the objects and relationships.

Appl/A : A Process Programming Language

The Arcadia research project is exploring several issues related to the development of process-oriented environments [Taylor88]. One research issue is the development of a suitable language for process programming. The language emerging from this research is Appl/A, a superset of Ada that enables the definition of relations among software objects. Ada was chosen as a base language because it seemed to support many of the capabilities envisioned as necessary in process programming. Ada was enhanced with a relation capability to provide a means to explicitly represent interconnections among complex objects.

Appl/A supports the definition of relations as sets of arbitrary tuples of software objects. It enables users to specify just how the various components of these tuples should be related to each other and how the consistency of these components can be verified and maintained.

To gain experience with Appl/A it has been used to write prototypes of several process programs. This experience has lead to the identification of several other necessary features. One such feature is the need for a type hierarchy. As an example, the different nodes of a requirement element may be of different types, but they also share some common features and they can be better modeled as being subtypes of a common parent type. Experience with Appl/A has lead to the impression that a strictly imperative, algorithmic language is not likely to be suitable as a process programming language. Some software activities seem best described with a declarative or rule based paradigm.

2.4 Selecting a Language

The current research into the software re-engineering process requires a suitable notation for recording and expressing the knowledge learned about software re-engineering. The current research focuses on two issues. The first issue is to identify the information that is fed into the re-engineering process and the information that is created and used within the process. The goal is to identify these information objects, determine their properties, and map their interrelationships. Further, the operations that are performed on these information objects are to be identified and described. The second issue is to identify the tasks that are done within the re-engineering process. Each task is associated with a collection of information. Tasks operate on information, and tasks interact with other tasks. All of this must be expressible within the process modeling language chosen.

To select a language for process modeling the features of the languages presented in the previous section were considered. For each language a small process model was developed to obtain a feel

for the expressness, clarity of expression, and adequateness of the language. The current research needs were compared with the capabilities of the various languages.

The major criterion used to evaluate the languages was the expressive power of the language. A process model is rich in detail and a language must provide suitable constructs for expressing process information. Two major concerns are the ability to express data information and task information.

A modeling language must be able to express the complex properties of the information (data) that is created, used, and manipulated within a process. A language must express the elements that form an information object and the type of information associated with each object. Properties of these elements and their relationships with other others must be expressible. It was found that the range of expressiveness of data information varied. Some languages focused on representing data information while others focused more on representing task information. A few languages, such as CML, provided good support for both.

A modeling language must be able to express the details of each task within a process and the relationships and dependencies between tasks. A process consists of many tasks. Further there is often a hierarchy among tasks, with some tasks representing major process activities and other tasks representing minor activities. Tasks have many details associated with them, such as the conditions under which a task can begin, who is responsible for the task, who is assigned to the task, the steps or procedures followed by the task, etc. Tasks may depend on human effort and/or be supported by tools. The modeling language must be able to express this information. All of the languages provide constructs for expressing task information. The range of expressiveness ranged from high-level abstractions of task sequencing to detailed representations of task processing steps.

The modeling language must support different levels of abstraction. Because the current research is examining the properties of the process as opposed to detailed processing steps, a high level of abstraction is required. However, the ability to lower the abstraction level in the future and stay within the same language would be convenient. After all, an important result of experience with process modeling is that a design of a process can reveal more about the process than a detailed process program. Most languages only provided one abstraction level. This level ranged from high-level descriptive modeling to detailed process programming.

Finally, the language chosen must be available for use. Almost all process modeling languages are experimental. While descriptions of these languages can be found, implementations or running systems are rare. Even complete language descriptions can be hard to obtain.

Given these considerations and after reviewing the example process models created with each language the language CML (Conceptual Modeling Language) developed for the TRIAD project was deemed the most suitable. This language has the expressive power needed and operates within a process modeling environment. Further, CML, which was originally developed and reported in several PhD theses, has been converted into a commercial product available from Universal Energy Systems.

However, while the authors liked the expressiveness of CML the best, it was decided that this language would not be used. First, the commercial system based on this language was not available to the authors. If CML were used to record the process model, there would be no tools for processing the model. Second, since CML is not widely known, its usefulness in conveying the process details

to others would be limited. Third, CML is a process programming language, and as such its level of abstraction is low since process executable information must be given. At best the current research would only result in a skeleton of a CML process model.

Since the current research is focusing on the information needs and major tasks of the re-engineering process it was decided to use a formal specification language. A formal specification language provides the expressiveness to convey *what* a process does and properties of information objects and operations on these objects, without forcing a description of *how* the operations and tasks are done. A formal language permits information to be expressed concisely and with precision. Further, reasoning about process properties is facilitated by a formal language. Finally, there are several well-known formal specification languages, thus a wider audience will be able to read and use the process model specification.

The language chosen to express the process model is the specification language **Z**. **Z** is a formal specification language based on typed set theory [Spivey89]. **Z** gains its simplicity and expressive power by using directly the well developed notations of mathematics. Originally developed for formal specification of software, the authors have found its expressive power suitable for specifying properties of the re-engineering process.

2.5 The Z Specification Language

Since **Z** is based on both set theory and predicate calculus it has the dual advantage of using a mature mathematical notation and of being widely accessible to those with some mathematical background. **Z** uses mathematical notation to describe in a precise way the properties which a software system must have, without describing how these properties are achieved. A **Z** specification is mathematical: the variables that appear in a specification range over mathematical objects. A **Z** specification expresses a mathematical model. This characteristic of **Z** specifications has proved useful in capturing and expressing the properties of the software re-engineering process.

Every mathematical expression which appears in a **Z** specification is given a type. This type determines a set known to contain the value of the expression. Each variable is given a type by its declaration. This use of mathematical data types to model the data in a system makes it possible to reason effectively about how a specified system will behave. **Z** uses predicate logic to describe abstractly the effect of operations on data.

Another main element of a **Z** specification is the use of *schemas*. A schema is used to specify a collection of variables, each with a type, and predicates on these variables that specify properties of the variables and relates the variables. A typical schema has the form:

```
┌─ Schema.Name ──────────────────────────────
│ Variable declarations
├──────────
│ Predicates
└────────────────────────────────────────────
```

This report assumes basic familiarity with **Z** and set theory. Two good references on **Z** are "The Z Notation: A Reference Manual" by Spivey [Spivey89] and "Specification Case Studies" by Hayes [Hayes87]. A summary of **Z** notation used in this report is given in Appendix A.

Chapter 3

Software Re-engineering Process

The realization that software development projects follow a definable sequence of phases was an important discovery within software engineering. The resulting Waterfall model was much talked about and many organizations adapted it for their use. As time went by, it was discovered that the Waterfall model did not fully model the interactions between software development phases. Indeed, it was even realized that alternative models of software development existed.

Like the Waterfall model for software development there is a traditional model for software re-engineering. The re-engineering model is based on three phases. It is a high-level, abstract model, devoid of most information about the process. Like the Waterfall model the traditional software re-engineering model focuses on the technical aspect of the process and fails to mention other aspects such as management or support.

One contribution of this research is an emphasis on studying the *complete* process of software re-engineering. A successful software re-engineering methodology or a useful environment for software re-engineering must be based on an understanding of the complete process, not just a portion of it. The emphasis on technical solutions evident in many research papers shows a lack of attention to the full range of concerns that occur during a software re-engineering project.

In the next section the traditional software re-engineering process model is presented. In section 3.2 a more complete model of the software re-engineering process is presented. Both process models are presented at a high level of abstraction. Chapter 4 gives a more detailed specification of one phase identified in the new process model.

3.1 Traditional Re-engineering Model

The goal of software re-engineering research is to develop the means to take an existing software system and generate from it a new system, that benefits from modern software engineering practices, and is functionally equivalent to the original system. In truth, a new system produced by re-engineering may not be functionally equivalent to the original system. Changes in the requirements for a system are often incorporated into the new system, introducing a difference between the functionality of the old and new systems. A fundamental assumption underlying the concept of

> The re-engineering of a software system produces a new form of the system that is better, in some way, than the original form.

In general, a software system is re-engineered because one or more of its properties are considered substandard. There are many properties associated with a software system such as the understandability of the source code, ease of maintenance, reliability, cost of maintenance, user satisfaction, ease of use, quality of documentation, size, etc. Empirical results support the assumption that re-engineering a system can indeed result in improvements [Sneed90].

Modern software engineering practices are based on the principle of **refinement**. Software is produced by a series of refinements where each refinement introduces a new detail into the system description. The choice of a particular refinement is guided by the development method in use. First, a specification is created. A design is created through a series of refinements on the specification. (Note this is not true of older development methodologies such as Yourdon-Constantine.) An implementation of the system is then created by another series of refinements on the design. The result is a software system that benefits from modern software development practices. In software engineering the principle of refinement is fundamental to software development, also called **forward engineering**.

A serious problem immediately arises when the application of modern software development practices to existing systems is considered. The problem is simply that the existing system has already been developed. A particular development technique expects to start with a specification or design for the existing system, instead of the source code. Unfortunately, most existing software systems are either undocumented or existing documentation is badly outdated. A second fundamental assumption underlying software re-engineering is that:

> The only valid, accurate, and up-to-date description of the functionality of an existing software system is the source code for that system. Therefore, any software re-engineering project must begin with the system source code.

The problem then is how can modern software development practices be applied?

The solution to this problem is the use of **abstraction**. Abstraction is applied to a software system by ignoring certain details of the system. This simplifies a task by ignoring details that are irrelevant to the current problem or task. Abstraction is the reverse of refinement. Where refinement adds successive layers of details, abstraction strips away details. The principle of abstraction is fundamental to **reverse engineering**. Chikofsky and Cross defined software reverse engineering as :

> The process of analyzing a software system to identify the system's components and their interrelationships and to create representations of the system in another form or at a higher level of abstraction [Chikofsky90].

Figure 3.1 shows the traditional diagram that contrasts forward and reverse engineering. Forward engineering begins with requirement analysis and specification, proceeds into design, and design is followed by implementation. Reverse engineering starts with an implementation and works backward.

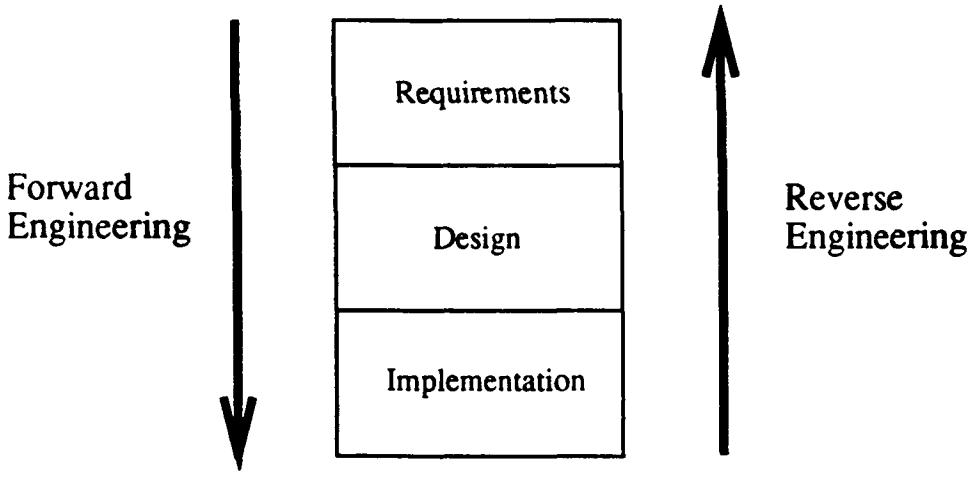


Figure 3.1: Forward and Reverse Engineering

The principles of abstraction and refinement embodied in reverse engineering and forward engineering respectively, form two pillars of software re-engineering. What is missing is the bridge between these pillars. This bridge is formed by the principle of **alteration**. Alteration is a change in the characteristics of a system. Alteration is done on information at some level of abstraction and includes the addition, deletion, and modification of information at that abstraction level. Within software re-engineering the principle of alteration is typically embodied by **re-design**. Re-design is the act of altering the design of the existing system to produce a design for the new system.

The traditional software re-engineering model can now be described. Figure 3.2 presents a diagram of this model. The traditional software re-engineering model begins with the source code for the existing system. First, the source code is analyzed. Starting with the analysis information a design for the existing system is constructed. This is reverse engineering. Implementation level details of the system are discarded and only information about the structure and functionality of the system are retained. The reverse engineering phase ends with a reconstructed design for the existing system.

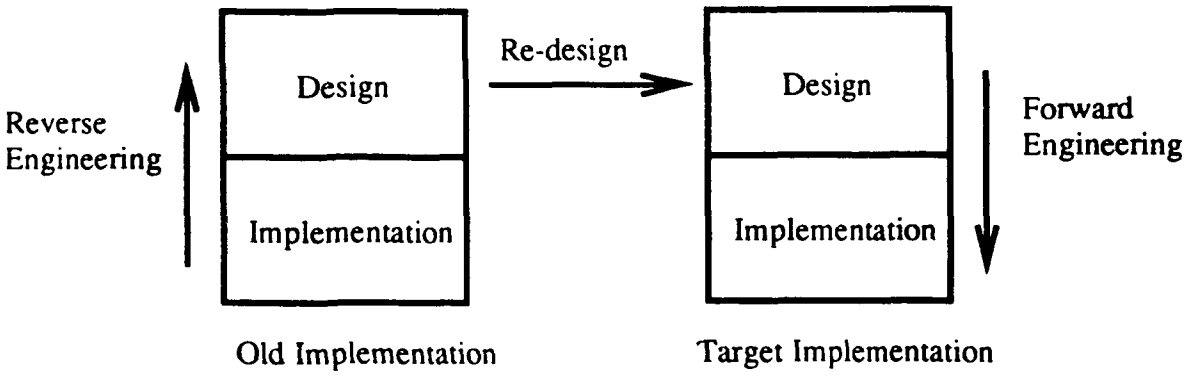


Figure 3.2: Traditional Software Re-engineering Model

The second phase, re-design, takes and alters the reconstructed design. Undesirable characteristics of the design, such as poor structure, are removed. The design may be changed to reflect

an object-oriented approach. New functionality may be woven into the design and existing system functionality may change. The result of the re-design phase is an altered design that reflects modern software design principles.

The third phase, re-implementation, takes the altered design and uses forward engineering techniques to create a new implementation of the system. This new system, referred to as the target system, reflects the benefits of using modern design and coding practices to implement software. The product of the re-implementation phase is the desired new system.

The traditional software re-engineering model shows three major phases. The sequencing of these phases is clearly shown. The main input and end-product of each phase is also given. However, there are many other inputs into the process and other outputs as well, that are not mentioned in the above description. This model is presented at a high level of abstraction. Lowering the abstraction level and the resultant increase in process detail would provide the answers to many questions about the software re-engineering process.

The problem with this model is that even if each phase was carefully expanded to the lowest level of process detail, there would remain unanswered questions about the re-engineering process. There would still be unexplained tasks that are part of any re-engineering project. The reason is that there are tasks within a re-engineering project that occur outside the domain or area of concern of the three specified phases. The traditional software re-engineering model focuses on the technical aspect of re-engineering and ignores the issues of managing and supporting a re-engineering project.

The key point to emphasize is that the traditional model of software re-engineering, which is shown in numerous articles on the subject, is incomplete. The traditional model ignores several aspects of the software re-engineering process. A useful process model must cover the full range of tasks that can occur during a re-engineering project. The failure of the model to cover the full range of tasks limits its usefulness for presenting, discussing, reasoning about, and improving the software re-engineering process.

3.2 New Re-engineering Model

A new model of the software re-engineering process is needed that includes the full range of software re-engineering tasks. The traditional software re-engineering model focuses on the technical aspects of re-engineering only. However, it must be realized that there is more to the re-engineering process than just the technical work.

3.2.1 Re-engineering Process Aspects

There are three aspects or areas of concern within the software re-engineering process. These aspects are:

Management : A software re-engineering project must be managed just like any other software project. Management tasks focus on project control. A re-engineering project requires planning. A project requires monitoring. A project requires direction and motivation to keep it on track and on schedule.

Technical : The heart of a re-engineering project is the technical work. It is the technical tasks that reconstruct the system design, change it, and re-implement it creating the target system. In addition, the target system must be tested. The documentation for the system must also be updated or replaced.

Support : The managerial and technical aspects of software re-engineering require support if they are to proceed smoothly. One useful support responsibility is configuration management, keeping track of all the information and objects used within a project. Quality assurance is a support task that monitors the output of tasks and ensures conformance with project standards. Monitoring of an on-going project and the collection of historical project information provide information to on-going managerial tasks.

These three aspects of the software re-engineering process are not phases. Management of a project does not stop and technical work begin at some point. Within each aspect is a sequence of tasks that occur through out a re-engineering project. These three aspects or areas of concern and the tasks that embody these concerns blend together to form a complete model of the software re-engineering process. Figure 3.3 represents these three aspects of the software re-engineering process.

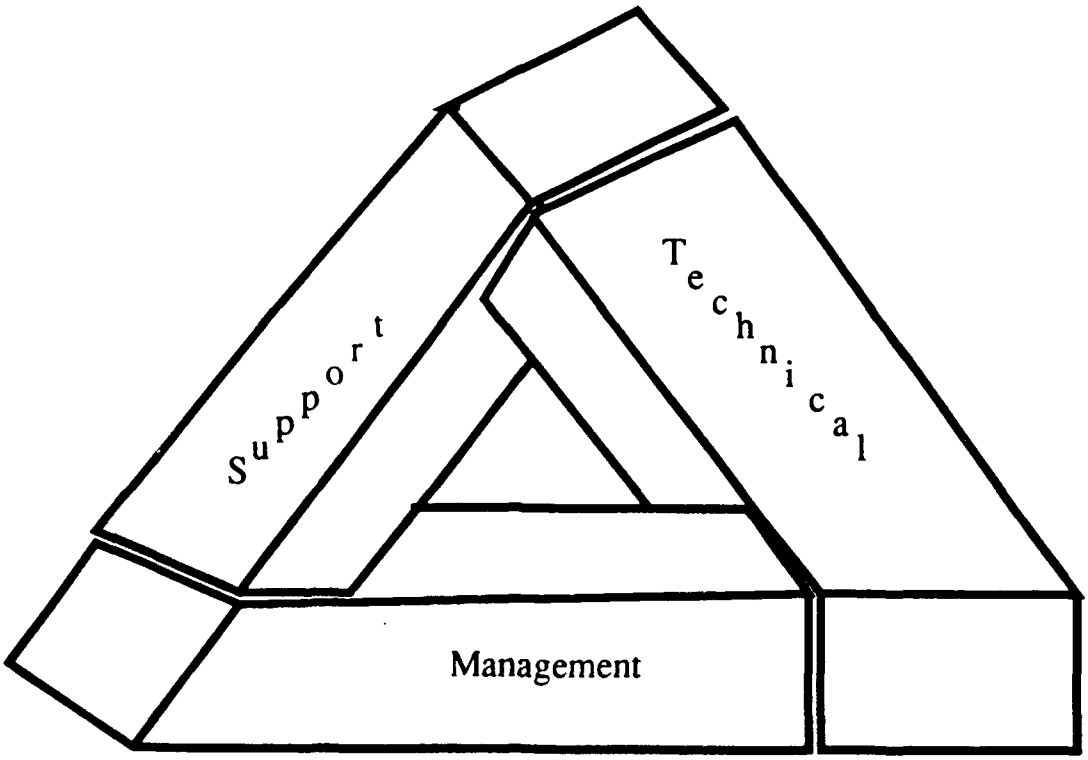


Figure 3.3: Aspects of the Software Re-engineering Process

Now that the aspects that must be incorporated into the process model are known, what does the software re-engineering process model look like? This question is addressed by examining each aspect in more detail.

Project Management

A software re-engineering project requires careful management. One important task of management is planning. A re-engineering project can be planned in a top-down manner. A general plan of the entire project is created first and then pieces of the plan progress to more specific levels of detail. Feedback is important and the on-going project must be carefully monitored. Project information must be analyzed and changes to the project plan made accordingly. The use of project tracking information is referred to as bottom-up input to the planning process.

Project management must deal with several sources of difficulties. These difficulties include people, technology, physical constraints, costs, organizational objectives, and external conflicts. Management must coordinate the many different tasks that occur during a project. The interfaces, dependencies and concurrences between the tasks and between the software being re-engineered must be taken into account. Finally, management must provide supervision over the many and varied types of individuals involved with a re-engineering effort.

Re-engineering tasks that represent management duties and concerns include:

Define Approach The project approach forms the project plan. The approach states how the existing software system will be re-engineered. The approach defines the methodology used in the project, the project phases, and the tasks within each phase. The approach determines if the system will be re-engineered in one lump sum, in a sequence of pieces, or only partially.

Estimation Project management must be able to estimate the effort, cost, duration, and staffing requirements for a re-engineering project. Estimates of cost are required if sufficient funds are to budgeted. Estimates of project duration are important for coordinating with the system users, and with other organizational projects. Estimates of staffing requirements are important for scheduling personnel that will be assigned to the project.

Define Organizational Structure Work within a re-engineering project is performed by teams. The organization of project teams can be guided by the project approach and the organization's (i.e. company's) philosophy. There can be a team for each project phase, or several teams can work in parallel doing similar tasks on different portions of the system.

Define Project Procedures and Standards Project management must determine the procedures that will be followed during the project. Procedures can specify how to conduct reviews, how to report errors, how to check out and in items from the configuration management system, etc. Project standards specify the acceptable format of products produced during the project.

Identify Resources Project management must determine resources need by the re-engineering project. Once the necessary resources are identified, they must be located, and arrangements made to have these resources available to the project. The project schedule can be influenced by the availability of necessary resources.

Plan System Transition Project management must plan how the target system will be phased into usage and the existing system phased out. Bringing the target system on-line signals the end of the re-engineering project, but it must be planned and coordinated in advance.

Scheduling A re-engineering project consists of many tasks. Each task must be scheduled. Tasks are given an expected starting date and stopping date. The project schedule represents a relationship between tasks. Some tasks can not begin until others are finished. Some tasks can be done concurrently. Some tasks will be repeated with different inputs.

Identify Tools There are many tools that can be used during a re-engineering project. Useful tools must be identified and their role within a project evaluated. Tools may need to be purchased, others may already exist in-house. Tools can reduce the project duration and improve system reliability, promote consistency, and free personnel from clerical work. Tools can be used to support all aspects of software re-engineering.

Define Acceptance Criteria Project management must negotiate with the client or system users and reach an agreement that states the criteria that must be satisfied for the target system to be accepted.

Conflict Resolution The on-going re-engineering project must be managed and any problems or conflicts that arise must be resolved by project management. Conflicts can be of a technical or non-technical nature. Technical conflicts are best resolved by technical staff, but occasionally management must step in and make a choice that solves a technical problem.

Project Authorization As the project moves from phase to phase and from major task to major task, project management must often give go-ahead permission for a task or phase to begin. This provides management with an opportunity to verify that a milestone has been reached and that the work done to date is satisfactory.

Personnel Management Project management is responsible for managing project personnel and maintaining project morale. People assigned to a project may quit the organization or ask to leave a project. A person may not do his/her job well, or may be disruptive to their team. Management must oversee the performance of project personnel and ensure that each person is working smoothly and satisfactorily towards the project completion.

The management aspect of the re-engineering process seeks to coordinate and control the many details of a re-engineering project. Project management does not concern itself with the details of the technical work, but must be familiar with the technical steps and what is involve.

Project Technical Work

It is the technical aspect of the process that does the re-engineering work. It is the technical tasks that manipulate the system source code, design, tests, and documentation. The traditional focus on the technical aspect of re-engineering is understandable. It here that most of the project work is done. Certainly, the most visible products of a re-engineering project are created by the technical tasks. The list of technical tasks includes the following:

Determine Motivations and Objectives This task focuses on two questions: Why is the system to be re-engineered? What does the project hope to achieve? The answer to the first question is a list of project motivations. Is the system difficult to maintain? What is happening to the system? The answer to the second question is a list of project objectives.

Analyze Environments If the system is to be ported to a new environment or if the existing system environment is to be upgraded or changed, the differences between the existing environment and the target environment must be stated. This analysis affects both the technical and managerial aspects of the project. A difference between environments can affect the changes made to the system. A new target environment may require staff training.

Collect Inventory The inventory task identifies and locates all source code files, system documents, documentation files, test data, test results, maintenance history records, etc. This information is turned over to the project librarian and becomes a support task problem, but the identification of all these objects requires technical knowledge of the system. The collection of all system pieces forms a baseline on which the reverse engineering task can begin.

Test Planning The target system must be thoroughly tested and this requires the development of a set of test inputs and outputs, known as a test plan. The existing system test plan can be evaluated and sections of it reused depending on the changes that will produce the target system. New tests for the target system must also be created.

Target System Testing This is the application of the test plan to the target system to verify that no errors can be detected. This is the same as testing done during software development.

Documentation Planning The existing system documentation must be collected and evaluated. Non-existent documents must be written. Badly outdated documents must either be replaced or rewritten. Other documents will also need to be evaluated and either replaced or rewritten to reflect the characteristics of the target system.

Create Documentation This is the work of re-documenting the system. The documentation plan guides the changes made to existing documents and directs the writing of new documents. The target system must have a complete, accurate, and up-to-date set of documentation.

Analyze Implementation To prepare for the design reconstruction work the existing system source code must be completely analyzed and its properties identified. The analysis records the implementation details of the existing system. A data dictionary of source code items is created.

Reconstruct Design This task uses the source code analysis information and reconstructs a design for the existing system. Implementation details are removed from the information and a design representation is created. The design records the system architecture and functionality. The reconstructed design may be expressed at several levels of abstraction.

Plan Design Changes The changes to be made to the reconstructed design are identified and formed into a plan. Information about the existing design and the envisioned target system must be studied. The resulting plan describes the changes to be made to the design.

Update Design The current system design is altered according to the re-design plan. Each change made to the design is the result of a design decision. A record of all changes must be kept.

Implement Design This task uses forward engineering techniques, takes the altered system design, and creates a new implementation of the system. Standard forward engineering techniques may need to be adapted to difficulties posed by re-engineering. For example, it is possible that only a portion of a system is re-engineered. Here, it will be necessary to merge the new target portion of the system with the retained and unmodified portions of the system.

Analyze New Source Code To document the target system source code, the new code must be analyzed. This is similar to the analysis done on the existing system source code.

Acceptance Testing This task verifies that the target system satisfies the acceptance criteria. Also the updated documentation must be verified, as well as converted file and database formats. Typically, the target system and existing system are executed in parallel, and a comparison of results is done.

System Transition This is the last task in the re-engineering process. This task controls how the target software is placed into production and how the existing system is phased out.

Project Support

Project support tasks are miscellaneous tasks that can not properly be considered as technical or managerial. These are tasks that handle details and duties that would interfere with the duties of other groups. Several supporting tasks are:

Configuration Management There are many physical entities to be identified, handled, monitored, and controlled. Each project has it own associated source code, listings, test data, test cases, test scenarios, test results, and documentation. Typically, a separate team is formed just to handle this task. At the head of this task is the project librarian, a person charged with tracking all the project items and their versions. Configuration management for software re-engineering is similar to that for software development. However, here a re-engineering project immediately starts with a large volume of information about the existing system, defined by the baseline inventory.

Quality Assurance The quality assurance task runs throughout the project. Typically, every product produced by a task is subject to a QA review. These reviews ensure that a product conforms to the project standards and is of satisfactory quality. The QA task also verifies that other project standards are adhered too.

Project Tracking The on-going project must be tracked and its progress monitored. Information about the project includes actual start and stop dates for tasks, the number of people assigned to a task, the hours worked by project personnel. Tracking information is analyzed by management to verify that the project is on schedule and to detect any possible problems. Tracking information is also recorded into a historical project database that can be used to review project estimates against actual values. This comparison can help fine-tune the estimation models used within an organization.

3.2.2 Re-engineering Project Phases

The new software re-engineering process model proposed in this report consists of seven phases. These phases cover not only the central focus of re-engineering, i.e. the production of the target system, but also project preparation, and the testing, documentation, and installation of the target system. Figure 3.4 presents a diagram of the software re-engineering process model. The software re-engineering process model contains these phases:

1. Analysis and Planning
2. Design Reconstruction
3. Re-design
4. Re-implementation
5. Target System Testing
6. Re-documentation
7. Acceptance and System Transition

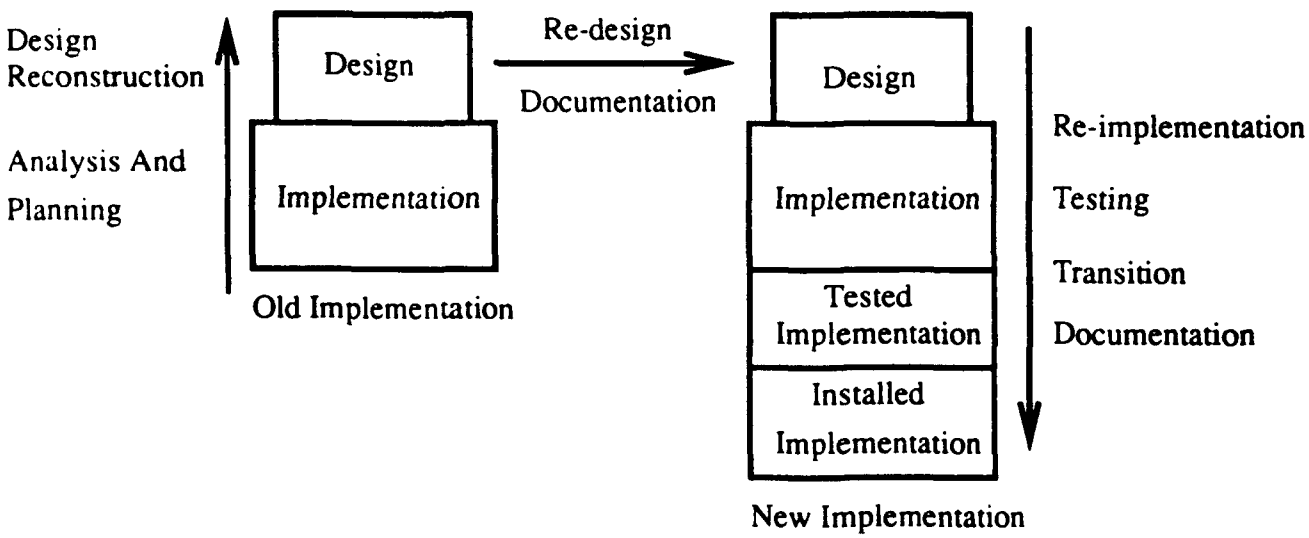


Figure 3.4: Software Re-engineering Process Model

Each phase of the process consists of a collection of tasks. Tasks may be performed sequentially or in parallel depending on the nature of the tasks and the organization of the project. Several of the tasks run throughout much of the process such as configuration management and project tracking. The tasks assigned to each phase are listed below.

Analysis and Planning

- Determine Motivations and Objectives
- Analyze Environments
- Collect Inventory
- Define Approach
- Plan System Replacement
- Define Acceptance Criteria
- Define Project Procedures and Standards
- Identify Resources
- Identify Tools
- Test Planning
- Documentation Planning
- Estimation
- Define Organizational Structure
- Scheduling

Design Reconstruction

- Analyze Implementation
- Reconstruct Design

Re-design

- Plan Design Changes
- Update Design

Re-implementation

- Implement Design

Target System Testing

- Target System Testing

Re-documentation

- Analyze New Source Code
- Create Documentation

Acceptance and System Transition

- Acceptance Testing
- System Transition

Not included in these lists are the miscellaneous tasks that either stand by themselves or do not fit into the area of concern embodied by a particular phase. These miscellaneous tasks are:

Miscellaneous Tasks

- Configuration Management
- Process Tracking
- Conflict Resolution
- Project Authorization
- Personnel Management
- Quality Assurance

The new software re-engineering process model presented in this chapter is a more complete model. The identification of several additional phases and the inclusion of many more tasks suggests the complexity of a re-engineering project. However, the model presented in this chapter is still a high-level abstraction of the re-engineering process. Many details of the process are have not been shown. In the next chapter the Analysis and Planning phase is covered. This coverage is at a lower level of abstraction and focuses on the information objects used and produced by the phase tasks. The tasks that occur during this phase are also discussed in more detail.

Chapter 4

Analysis and Planning Phase

The analysis and planning phase is the first phase of the software re-engineering process. If planning and organization are the keys to a successful project then this is the key phase. There may be an urge to hurry up and just produce a cursory project plan or to skip this phase all together so that the "real" project work can begin. That is a mistake, only by careful planning can a re-engineering project be conducted smoothly and with a minimum of unexpected problems.

The analysis portion of this phase focuses on technical aspects of a project. Analysis addresses three issues. The first issue is to define the current state of the existing system. What are its shortcomings? How serious are these shortcomings? Because the re-engineering work will begin with the existing system source code, the properties of the existing system must be understood.

The second issue is to define why the existing system is to be re-engineered. Is the project prompted by dissatisfaction with one or more aspects of the system? Are there external influences involved? Is the system to be ported to a new computer? Does the implementation language need to be changed due to some organizational directive? Has the existing environment been upgraded and the system is to be changed to take advantage of the new environment? Is the motivation a mix of dissatisfaction and external influences. The project motivations must be understood since they directly influence the project objectives.

The third issue is to specify the characteristics of the desired target system. This involves two aspects of the system. First, are there changes to the system requirements? Will the functionality of the target system differ from that offered by the existing system? If yes, then the new requirements must be specified, just like a software development project. Are there changes to existing requirements? Are any existing requirements to be deleted? If the answer is yes to either question, the affected requirements must be noted. If may be difficult to exactly state which existing requirements are to be changed or deleted, since the system may have no valid requirement specification. The second aspect involves changes to the characteristics of the system. Is the system design to be changed? Is the system implementation to be changed? Changes of this nature can include translation to another language, cleaning up of program logic, the introduction of modularized or object-oriented design principles, etc. The desired characteristics of the target system must be understood to plan the work necessary to create the target system.

The planning portion of this phase focuses on management aspects of a project. To conduct

a successful project it is necessary to understand the scope of the work, the resources required, the tasks to be done, the milestones to be tracked, effort to be expended, and the schedule to be followed. Project planning focuses on research and estimation. Research is done to define the scope of the work to be achieved. Estimation is done to predict the project effort, duration, and cost. Estimates are used to guide project planning.

The remainder of this chapter is devoted to a formal specification of the analysis and planning phase. The specification is not complete, in that all aspects of the phase are not specified. The specification focuses on two aspects of the phase: information and operations on information.

What information is used and created during this phase? The re-engineering process centers on information. Information flows into the process, is manipulated, created, and produced as a product of the process. Each item of information must be identified. Each property specified, each relationship with other information elements clarified. The role of each element of information within the process must be stated. To understand the process it is necessary to understand the information on which the process operates.

What operations are performed on the information used in this phase? Information is data. Elements of information are grouped together and treated as objects. Information is read, created, changed, and deleted. It is combined to form new information. Operations are done on information to achieve task goals, phase goals, and project goals. The operations that can be done on each element of information must be specified. Understanding these operations clarifies what can and can not be done within the process.

In addition to these two issues, the specification focuses on tasks that occur within the phase. Descriptions of tasks are given, but formal properties and interrelationships between tasks are not yet provided. One last issue that is also addressed is the role of knowledge within the process. Humans apply knowledge to their work. Often the role of knowledge is ignored, but to completely specify the re-engineering process, the role of knowledge within the process must be considered.

4.1 Types of Knowledge

The use of knowledge permeates the entire software re-engineering process. However, the role of knowledge within the process is often ignored or over looked. This becomes apparent when process tasks are considered. For example, one task within the analysis and planning phase is to produce a list of project objectives. Where do these objectives come from? What input or source of information can be given that accounts for this output (the list of objectives) from the task? These objectives derive from a human's understanding of the existing system's status, its environment, and the needs of the system's users. The list of objectives is not created by a computerized tool that analyzes information stored in a computer. The list is created by a human using several sources and areas of knowledge and then *recorded* using a computerized tool onto a computer file.

This example emphasizes an important point in process modeling. Many useful tools serve clerical roles. Humans do the work and tools are used to record the result of human labor. As our knowledge and understanding of parts of the process deepen better tools that take on more of the burden of the work will be created. Knowledge can be embedded into tools that can then do work that currently is a human responsibility.

The issue of knowledge and its role in the process is important. There are many different knowledge domains that are used in the software re-engineering process. Currently, it is not possible to fully define each knowledge domain. Rather than enter into a detailed specification of knowledge, its structure, and the relationships between different groups of knowledge, we will specify knowledge at a very abstract level.

The set of all knowledge is denoted by KNOWLEDGE. The structure of this set will be left unspecified. This may be unsatisfactory, but it avoids a can of worms that would side track this discussion. Several subsets of KNOWLEDGE can be specified. These subsets serve to group together knowledge related to a particular domain. Again, intuitive definitions of these sets will be given. Each knowledge domain has the specification:

$$domain \subset KNOWLEDGE$$

The knowledge domains of interest are explained below. There are possibly other knowledge domains used in software re-engineering that are not listed. As our understanding of the software re-engineering process grows it is likely that other necessary domains will be identified.

Application Domain Knowledge This is knowledge about the characteristics and properties of the problem area that the system addresses. A software system encodes a process for solving a problem or performing some task. To understand why a system does what it does, it is necessary to understand the problem area in which the system operates.

Environment Knowledge This is knowledge about the computer environment in which the system executes. It is knowledge about the computer hardware, the operating system, and other systems and devices with which the existing system interacts. Environment knowledge includes knowledge about any changes or planned changes to the existing system environment or the target environment that the target system will operate in. This knowledge includes an understanding of the computing environment at the organization or the client site.

Existing System Knowledge This is knowledge about the characteristics and properties of the existing system. This is an understanding of how to use the system, how it works, the source code, the system history, where the source files are located, the documentation, and the shortcomings and strengths of the system.

Management Knowledge This is knowledge about how to manage a project. This is knowledge about scheduling, estimation methods, dealing with personnel, understanding potential management pitfalls and how to avoid them, etc. This is not specific knowledge about managing re-engineering projects, though that is included. It is a broad understanding about how to manage software projects.

Organization Knowledge This is knowledge about the organization that will conduct the software re-engineering project. Typically, only a subset of the organizational resources (i.e people and equipment) are dedicated to a project, but an understanding of the entire organization, its structure, its policies and procedures is required.

Programming Language Knowledge This is knowledge about the languages used to implement the existing and target systems. It is an understanding of the language principles, the syntax,

Software Engineering Knowledge This is knowledge about software engineering principles and practices. Forward engineering is a component of the re-engineering process so general software engineering techniques must be known.

Software Re-engineering Knowledge This is knowledge about how to re-engineer software, techniques, methods, and methodologies. It is the understanding about how re-engineering work is accomplished. This includes knowledge about necessary tasks, techniques, methods, pitfalls and how to avoid them or get out of them, etc.

Tool Knowledge This is an understanding of tools that are available either in-house or commercially. It is knowledge about how tools can be used, the problems they address, and their suitableness for any given task. Tools can play an important role within a project and that role must be understood if the tools are to be used effectively.

4.2 Information and Operations

This section presents a formal specification of the information used within the Analysis and Planning phase of the software re-engineering process. Operations on information items are also specified. For each information object, a description of the information and its role in the process is given, followed by a specification of the information. Finally, operations on the information are described and specified.

Error conditions and error handling are not currently covered in the specification. While it is possible to specify errors, it was felt that the extra detail would detract from the essential points.

4.2.1 Basic Sets

There are several sets that are used through the specification. These sets are defined here. The index provided with this report shows where each defined item of the specification is declared and used. The reader may wish to refer to the index while reading the specification.

Dates are used throughout the re-engineering process. A project has a start date and stop date. Tasks and phases have estimated start and stop dates, as well as actual start and stop dates. Employees record hours worked on particular dates. The set of all valid dates is denoted by the set DATE.

$$DATE \mathrel{\hat{=}} \textit{The set of all valid dates}$$

We let $\epsilon \in DATE$ denote an unassigned date.

Names are used throughout the re-engineering process. Programs have names, files have names, personnel have names, tasks have names, etc. The set of all possible names is denoted by the set NAME.

$$NAME \mathrel{\hat{=}} \textit{The set of all names}$$

We let $\epsilon \in NAME$ denote an unassigned name.

Many items of information are *labeled.* Individual elements of a requirement specification have labels, steps in a procedure are labeled, elements of any list are typically labeled. The concept of a label differs from the concept of a name. An item that is labeled typically is part of a sequence of items or a collection of items that are somehow ordered. Whereas, a name denotes individuality within a set of related items. The set of all possible labels is denoted by the set LABEL.

$LABEL \mathrel{\widehat{=}}$ *The set of all labels*

Many times throughout a process it is necessary to specify that some action can not happen until a set of conditions has been satisfied. Or that some task will satisfy a set of conditions. A condition may specify that a certain data object must exist or that another process step has been completed, or that a human has permission to begin the action associated with a step. We let CONDITION denote the set of all possible conditions. (For now, we ignore the fact that a condition will most likely be a collect of conditions connected by logical connectives.)

$CONDITION \mathrel{\widehat{=}}$ *The set of all possible conditions*

4.2.2 Property Lists

Many of the information objects that are manipulated during a re-engineering project have an associated collection of *properties.* These properties differ from object to object. For example, a source code file has a name, a size, a list of functions defined in it, a programming language associated with it, etc. These are properties of source code files.

A property typically has a name and a value. The set of all possible properties is denoted by the set PROPERTY. The set of all possible property names is denoted by the set PROP_NAME. The set of all possible property values is denoted by the set VALUE.

$PROPERTY \mathrel{\widehat{=}}$ *The set of all properties*
$VALUE \mathrel{\widehat{=}}$ *The set of all property values*

$PROP_NAME \mathrel{\widehat{=}}$ *The set of all property names*

$PROP_NAME \subset NAME$

A list of properties is modeled as a partial function from property names to values:

$PROP_LIST : PROP_NAME \rightarrow VALUE$

A initial property list is empty. This is denoted by the following constant:

$INITIAL_PROP_LIST = \varnothing$

Property List Operations

- Add-property
- Update-property
- Delete-property
- Get-property
- List-properties

To denote that a property list is changed by an operation the following notation is used:

$$\Delta PROP_LIST \mathrel{\widehat{=}} [\, pl, pl' : PROP_LIST \,]$$

To specify that a property list is not changed by an operation the following notation is used:

$$\Xi PROP_LIST \mathrel{\widehat{=}} [\, pl, pl' : PROP_LIST \mid pl' = pl \,]$$

The operation *Add-property* is used to add a new property to an existing property list.

Add-property

$$\begin{array}{l}
\Delta PROP_LIST \\
p? : PROP_NAME \\
v? : VALUE \\
\hline
p? \notin \operatorname{dom} pl \\
pl' = pl \cup \{p? \mapsto v?\}
\end{array}$$

The operation *Update-property* changes the value associated with a property.

Update-property

$$\begin{array}{l}
\Delta PROP_LIST \\
p? : PROP_LIST \\
v? : VALUE \\
\hline
p? \in \operatorname{dom} pl \\
pl' = pl \oplus \{p? \mapsto v?\}
\end{array}$$

The operation *Delete-property* removes a property from a property list.

Delete-property

$$\begin{array}{l}
\Delta PROP_LIST \\
p? : PROP \\
\hline
p? \in \operatorname{dom} pl \\
pl' = \{p?\} \lhd pl
\end{array}$$

The operation *Get-property* retrieves the value associated with a property.

Get-property

$\equiv PROP_LIST$
$p? : PROP_NAME$
$v! : VALUE$

$p? \in \mathrm{dom}\, pl$

$v! = pl(p?)$

The operation *List-properties* retrieves a list of all the properties and their values.

List-properties

$\equiv PROP_LIST$
$list! : \mathbf{P}\{PROP_NAME \times VALUE\}$

$list! = \{ p : PROP_NAME;\ v : VALUE \mid pl(p) = v \}$

4.2.3 Motivations and Objectives

The *motivations* for re-engineering a software system often reflect problems with the existing system. There are a variety of reasons to re-engineer a system. The primary reason is often to improve the system maintainability in an effort to reduce maintenance costs or shorten the time required to change the system. The motivations for a project cause the project to happen. Motivations are typically problems or events that must be addressed. Possible re-engineering project motivations include:

- The existing system is too costly to maintain.
- It takes too long to make changes to the existing system.
- The existing system (or part of it) must be rewritten in another programming language.
- The existing system is unreliable.
- The system must be ported to a new operating environment. This refers to adapting the system to new hardware or operating system.
- The system must adapt to changes in its operating environment. This may be caused by a new release of an operating system, compiler, or hardware.
- The documentation for the system is not complete, non-existent, out-dated, or badly inaccurate.
- The system uses old home-grown utilities.
- The system does not make use of newer programming language features.

The concept of re-engineering project motivations is formalized by defining the set of all possible motivations to re-engineer a software system. This is denoted by the set:

$$MOTIVATION \mathrel{\widehat{=}} \textit{The set of all possible re-engineering motivations}$$

Project *objectives* are typically the response to project motivations, however, that is not always true. Once the decision is made to re-engineer a system other objectives are often added. Other objectives, such as a desire to make changes to the system requirements, are often combined into a re-engineering project. (The old *while we're at it, why don't we also do this...* cliche.) Objectives are project goals. Objectives specify what the project is to accomplish. Possible re-engineering project objectives include:

- Improve the maintainability of the software system.
- Translate the system (or part of it) into another programming language.
- Improve the reliability of the system.
- Port the system. (System conversion.)
- Adjust the system to operate within the new environment.
- Change the functionality provided by the system.
- Re-document the system. Create new documents when a required document is non-existent. Update or replace existing documents.
- Replace home-grown utilities with environment supplied utilities. For example, replace locally written file manipulating routines with standard operating system calls.
- Make better use of programming language capabilities and features.

The concept of re-engineering project objectives is formalized by defining the set of all possible objectives for a re-engineering project. This is denoted by the set:

$$OBJECTIVE \mathrel{\widehat{=}} \textit{The set of all possible re-engineering objectives}$$

There are typically several project motivations and objectives, so it is helpful if each is labeled. This simplifies the problem of referring to a particular project motivation or objective.

The list of project motivations and objectives provide a high-level description of what a re-engineering project is to accomplish (goals) and the reasons for the project. These two lists are grouped together to form the project definition. The project definition may also be called the project overview. The project definition is formalized as:

$$\begin{array}{|l}
\hline
\textit{DEFINITION} \\
\textit{reasons} : LABEL \rightarrowtail\!\!\!\!\rightarrow MOTIVATION \\
\textit{goals} : LABEL \rightarrowtail\!\!\!\!\rightarrow OBJECTIVE \\
\hline
\end{array}$$

Notice that each labeled list is formalized as a one-to-one partial function. This captures the idea that a particular motivation or objective only occurs once, and has an unique label.

The initial value for the project definition is given by the constant INIT_DEFINITION.

$$
\begin{array}{|l}
\hline
\textit{INIT_DEFINITION} \\
\textit{DEFINITION} \\
\hline
\textit{reasons} = \emptyset \\
\textit{goals} = \emptyset \\
\hline
\end{array}
$$

Motivations and Objectives Operations

The project definition is manipulated by several operations. These operations are :

- Add-reason
- Add-goal
- Delete-reason
- Delete-goal
- Get-reason
- Get-goal
- List-reasons
- List-goals

The need to add new reasons and goals is clear. The need to delete a reason or goal is necessary due to changing views of the project. The initial list of reasons and goals is likely to change as the project scope is explored in more depth.

To specify that the project definition is changed by an operation the following notation is used :

$$\Delta DEFINITION \mathrel{\widehat{=}} [\, DEFINITION, DEFINITION' \,]$$

To denote that the project definition is not changed by an operation the following notation is used:

$$\Xi DEFINITION \mathrel{\widehat{=}} [\, DEFINITION, DEFINITION' \mid DEFINITION' = DEFINITION \,]$$

The operation *Add-reason* is used to add a new reason to the list of reasons for re-engineering a system.

Add-reason

$\Delta DEFINITION$
$m? : MOTIVATION$
$l? : LABEL$

$l? \notin \mathrm{dom}\ reasons$
$reasons' = reasons \cup \{ l? \mapsto m? \}$

$goals' = goals$

The operation *Delete-reason* is used to remove the specified reason from the list of reasons.

Delete-reason

$\Delta DEFINITION$
$l? : LABEL$

$l? \in \mathrm{dom}\ reasons$
$reasons' = \{l?\} \lhd reasons$

$goals' = goals$

The operation *Get-reason* is used to retrieve a reason for re-engineering a system from the project definition, given the label for the reason.

Get-reason

$\Xi\ DEFINITION$
$l? : LABEL$
$m! : MOTIVATION$

$m! = reasons(l?)$

The operation *List-reasons* is used to retrieve a list of all the reasons for re-engineering a system.

List-reasons

$\Xi\ DEFINITION$
$l? : LABEL$
$m! : \mathbf{P}\{ LABEL \times MOTIVATION \}$

$m! = \{ l : LABEL;\ m : MOTIVATION \mid reasons(l) = m \}$

The operations *Add-goal*, *Delete-goal*, *Get-goal*, and *List-goals* are similarly defined.

4.2.4 Operating Environments

A software system is assembled and operated within an environment. This environment is defined by the operating system, the computer architecture, the compiler, other systems that interact with

the existing system, devices used by the system, vendor supplied libraries of routines, etc. If any of these elements are changed, it may be necessary or desirable to adapt the existing system to conform with the new environment.

A change in the environment can be caused by an upgrade in the operating system or the compiler. It may be a change in the interface to another system. The existing system may be ported to a new computer with a similar or dissimilar operating system and different hardware architecture. If the target environment is not the same as the existing environment, it is necessary to analyze the two environments, detect differences between them, and determine the possible effects of these differences on the existing system. Environmental differences that are important to a re-engineering project are those differences that can affect the way the system operates.

Information about both the existing environment and the target environment must be collected. Environment information includes the type of computer, operating system, and its release level, vendor supplied utilities, utility interfaces, word size, byte ordering, character representation, etc.

Information about a particular environment is modeled by a property list. There are two environments that must be modeled: the existing and target environments. Information about these environments is formalized by *existing_env* and *target_env*, where:

$$existing_env : PROP_LIST$$
$$target_env : PROP_LIST$$

Once each environment has been analyzed the differences between the two environments must be noted. The differences are modeled by *diffs*, where *diffs* lists properties of the two environments that differ. For example, both environments may provide a system call to open a file, but the parameters used by the call differs between the two environments.

$$diffs : PROP_NAME \rightarrow VALUE \times VALUE$$

The existing environment features that are missing in the target environment must be noted. Features of the target environment that are not present in the existing environment must also be noted. These two sets of information are modeled by *lacking* and *new_features*.

$$lacking : \mathbf{P}\ PROP_NAME$$
$$new_features : \mathbf{P}\ PROP_NAME$$

The set *lacking* lists properties and features of the existing environment that are not found in the target environment. The set *new_features* lists properties and features of the target environment that are not found in the existing environment.

Finally, we want to know about the differences that are important. This is information is modeled by:

$$concerns : PROP_LIST$$

The partial function *concerns* does not list properties of the environments, but concerns (properties) and statements about those concerns (values). The concerns recorded refer to differences between the two environments that may influence the system.

All the above information is collected together into a schema that represents the environment analysis. This schema also states the relationships between these information items.

$$
\begin{array}{l}
\textit{ENVIRONMENTS} \\
\hline
\textit{existing_env} : \textit{PROP_LIST} \\
\textit{target_env} : \textit{PROP_LIST} \\
\textit{diffs} : \textit{PROP_NAME} \rightarrow\!\!\!\!\!+ \textit{VALUE} \times \textit{VALUE} \\
\textit{lacking} : \mathbb{P}\ \textit{PROP_NAME} \\
\textit{new_features} : \mathbb{P}\ \textit{PROP_NAME} \\
\textit{concerns} : \textit{PROP_LIST} \\
\hline
\textit{diffs} = \{ p : \textit{PROP_NAME};\ \ v_1, v_2 : \textit{VALUE} \mid \\
\qquad\qquad p \in \operatorname{dom} \textit{existing_env} \ \wedge\ p \in \operatorname{dom} \textit{target_env} \ \wedge \\
\qquad\qquad v_1 = \textit{existing_env}(p) \ \wedge\ v_2 = \textit{target_env}(p) \ \wedge\ v_1 \neq v_2 \ \} \\
\\
\textit{lacking} \subseteq \operatorname{dom} \textit{existing_env} \\
\textit{lacking} \cap \operatorname{dom} \textit{target_env} = \varnothing \\
\\
\textit{new_features} \subseteq \operatorname{dom} \textit{target_env} \\
\textit{new_features} \cap \operatorname{dom} \textit{existing_env} = \varnothing \\
\hline
\end{array}
$$

Initially, when the Analysis and Planning phase begins no environment information is recorded. The initial environment information is denoted by the constant INIT_ENVIRONMENTS.

$$
\begin{array}{l}
\textit{INIT_ENVIRONMENTS} \\
\hline
\textit{ENVIRONMENTS} \\
\hline
\textit{existing_env} = \varnothing \\
\textit{target_env} = \varnothing \\
\textit{diffs} = \varnothing \\
\textit{lacking} = \varnothing \\
\textit{new_features} = \varnothing \\
\textit{concerns} = \varnothing \\
\hline
\end{array}
$$

Operating Environment Operations

Several operations on environment information can be defined. Because most of the information is stored as property lists the property list operations can be used. The operations on environment information are:

- Add existing-env-prop
- Delete-existing-env-prop
- Update-existing-env-prop
- Get-existing-env-prop

- List-existing-env-props
- Add-target-env-prop
- Delete-target-env-prop
- Update-target-env-prop
- Get-target-env-prop
- List-target-env-props
- List-differences
- List-lacking
- List-new-features
- Note-concern
- Delete-concern
- Update-concern
- Get-concern
- List-concerns

Notice that there are no operations that set the values of *diffs*, *lacking*, or *new_features*. This information can be determined automatically as the specification shows.

To denote that the ENVIRONMENTS information can be changed by an operation the following notation is used.

$$\Delta ENVIRONMENTS \mathrel{\widehat{=}} [\, ENVIRONMENTS,\ ENVIRONMENTS' \,]$$

To denote that the ENVIRONMENTS information is not changed by an operation the following notation is used.

$$\Xi ENVIRONMENTS \mathrel{\widehat{=}} [\, ENVIRONMENTS,\ ENVIRONMENTS' \mid ENVIRONMENTS' = ENVIRONMENTS \,]$$

The operation *Add-existing-env-prop* is used to add a new property to the list of information about the existing environment.

Add-existing-env-prop

$\Delta ENVIRONMENTS$
$p? : PROP_NAME$
$v? : VALUE$

Add-property[$existing_env/pl,\ existing_env'/pl'$]

$target_env' = target_env$
$diffs' = diffs$
$lacking' = lacking$
$new_features' = new_features$
$concerns' = concerns$

The operation *Delete-existing-env-prop* is used to remove a property from the list of information about the existing environment.

Delete-existing-env-prop

$\Delta ENVIRONMENTS$
$p? : PROP_NAME$

Delete-property[$existing_env/pl,\ existing_env'/pl'$]

$target_env' = target_env$
$diffs' = diffs$
$lacking' = lacking$
$new_features' = new_features$
$concerns' = concerns$

The operation *Update-existing-env-prop* is used to change the information about a property of the existing environment.

Update-existing-env-prop

$\Delta ENVIRONMENTS$
$p? : PROP_NAME$
$v? : VALUE$

Update-property[$existing_env/pl,\ existing_env'/pl'$]

$target_env' = target_env$
$diffs' = diffs$
$lacking' = lacking$
$new_features' = new_features$
$concerns' = concerns$

The operation *Get-existing-env-prop* is used to retrieve an item of information about the existing environment.

Add-existing-env-prop

Ξ *ENVIRONMENTS*
$p? : PROP_NAME$
$v! : VALUE$

Get-property[*existing_env/pl*, *existing_env'/pl'*]

The operation *List-existing-env-props* is used to report all the information collected about the existing environment.

List-existing-env-prop

Ξ *ENVIRONMENTS*
$list! : \mathbf{P}\{ PROP_NAME \times VALUE \}$

List-properties[*existing_env/pl*, *existing_env'/pl'*]

The operations on target environment properties, *Add-target-env-prop*, *Delete-target-env-prop*, *Update-target-env-prop*, *Get-target-env-prop*, and *List-target-env-props* are defined similarly to the above operations.

The operation *List-differences* is used to obtain a list of all the environment differences.

List-differences

Ξ *ENVIRONMENTS*
$list! : PROP_NAME \times VALUE \times VALUE$

$list! = \{ p : PROP_NAME;\ v_1, v_2 : VALUE \mid diffs(p) = (v_1, v_2) \}$

The operation *List-lacking* is used to obtain a list of all the properties of the existing environment that are lacking in the target environment. The absence of a feature used by the existing system can increase the complexity of a project, since it usually requires that a similar feature be developed in the target environment.

list-lacking

Ξ *ENVIRONMENTS*
$list! : \mathbf{P}\ PROP_NAME$

$list! = lacking$

The operation *List-new-features* is used to obtain a list of all the properties of the target environment that are lacking in the existing environment. There may be features in the target environment that the system should be adapted to use. This operation is specified similarly to the operation *List-lacking*.

The operation *Note-concern* is used to note a property of the two environments that may be a concern to the project. The information recorded is not the property in question, but a concern about the property difference and information about that concern.

Note-concern

$\Delta ENVIRONMENTS$
$c? : PROP_NAME$
$e? : VALUE$

$Update - property[concerns/pl,\ concerns'/pl,\ c?/p?,\ e?/v?]$

$existing_env' = existing_env$
$target_env' = target_env$
$diffs' = diffs$
$lacking' = lacking$
$new_features' = new_features$

The operation *Delete-concern* is used to remove a concern that has previously been recorded. This operation may be necessary as the understanding of the project, system, and environments develops.

Delete-concern

$\Delta ENVIRONMENTS$
$c? : PROP_NAME$

$Delete - property[concerns/pl,\ concerns'/pl,\ c?/p?]$

$existing_env' = existing_env$
$target_env' = target_env$
$diffs' = diffs$
$lacking' = lacking$
$new_features' = new_features$

The operation *List-concerns* is used to obtain a list of all the concerns about environment differences that have been recorded.

List-concerns

$\Xi\ ENVIRONMENTS$
$list! : \mathbb{P}\{\ PROP_NAME \times VALUE\ \}$

$List - properties[concerns/pl, concerns'/pl']$

4.2.5 Inventory

The inventory lists the objects that are to be re-engineered or used during a re-engineering project. These objects must be identified and properties of these objects must be collected. The objects to be collected vary based on how the system is viewed. There are at least two viewpoints: the logical view and the physical view. Both logical and physical information must be recorded, as they will be used throughout the project.

The logical viewpoint sees the system as composed of objects such as programs, JCL commands, data stores, databases, and routines taken from libraries. There are also supporting objects such

as documents. Documents can be user manuals, reference guides, system requirement documents, design documents, maintenance history records, etc. Other supporting objects include test plans which consist of test cases, test input data, and logged test output data.

The physical viewpoint sees that system objects are stored in files. Files have properties such as name, and size, as well as record layouts for datafiles, programming language type for source code files, etc. Documents may be stored in files or perhaps only a hardcopy exists. Test plans are typically documents, but test cases are typically input data files and logged output data files.

To formally represent the inventory, we begin by naming the set of all possible system objects. This set is denoted by OBJECT.

$$OBJECT \mathrel{\widehat{=}} \textit{The set of all possible system objects}$$

Every object has a name or an id. To simplify things at this point we assume that all object names are unique. The set of all possible object names is denoted by OBJ_NAME.

$$\begin{array}{|l} OBJ_NAME \mathrel{\widehat{=}} \textit{The set of all possible object names} \\ \hline OBJ_NAME \subset NAME \end{array}$$

Each object has a set of properties associated with it. One property that every object has is a *type*. The *type* reflects the viewpoint on the object. Possible types include **program**, **JCL script**, **database**, **test plan**, **source code file**, **data file**, **document file**, **test data file**, **library file**, **design document**, etc. Furthermore, each object of the same type must have the same properties, i.e the same information should be collected for similar objects. This is specified by the set OBJECT_PROPERTIES which denotes the set of predefined property lists for different types of objects.

$$\begin{array}{|l} OBJECT_PROPERTIES \mathrel{\widehat{=}} \textit{The set all predefined object property lists} \\ \hline OBJECT_PROPERTIES \subset PROP_LISTS \\ \nexists\, ol : OBJECT_PROPERTIES \bullet ol = \varnothing \\ \forall\, ol : OBJECT_PROPERTIES \bullet \mathbf{type} \in \operatorname{dom} ol \end{array}$$

The predicates state that the property list for an object can not be empty and must have a property called **type**.

We define a schema INVENTORY which characterizes the system inventory information.

$$\begin{array}{|l} \textit{INVENTORY} \\ \hline inv_list : \mathbb{F}\, OBJ_NAME \\ inv_info : OBJ_NAME \rightarrow PROP_LIST \\ \hline inv_list = \operatorname{dom} inv_info \\ \forall\, obj : OBJ_NAME;\ obj2 : OBJ_NAME \bullet \\ \quad inv_info(obj)(\mathbf{type}) = inv_info(obj2)(\mathbf{type}) \Rightarrow \\ \qquad \operatorname{dom} inv_info(obj) = \operatorname{dom} inv_info(obj2) \end{array}$$

Notice that *inv_list* is defined as a finite set. This states that the number of objects associated with a system is finite.

At the start of a project, no inventory information exists. The initial inventory information is denoted by INIT_INVENTORY.

INIT_INVENTORY

INVENTORY

$inv_list = \emptyset$

$inv_info = \emptyset$

Inventory Operations

The system inventory is manipulated by several operations. These operations are:

- Add-Inv-Object
- Del-Inv-Object
- Update-Inv-Obj-Prop
- Look-Inv-Obj-Prop
- List-Inv-Obj-Props
- List-Inv-Objects

Notice there are no operations to add or delete properties for an object. The type of an object determines the properties that object possesses.

To denote that the inventory is changed by an operation the following notation is used :

$$\Delta INVENTORY \mathrel{\widehat{=}} [\, INVENTORY,\ INVENTORY' \,]$$

To denote that the inventory is not changed by an operation the following notation is used :

$$\Xi INVENTORY \mathrel{\widehat{=}} [\, INVENTORY,\ INVENTORY' \mid INVENTORY' = INVENTORY \,]$$

To add a new object to the inventory the *Add-Inv-Object* operation is used. By using this operation, a new object is added and its property list is initialized.

Add-Inv-Object

$\Delta INVENTORY$

$object? : OBJ_NAME$

$type? : VALUE$

$object? \notin inv_list$

$\exists\, empty_type_list : OBJECT_PROPERTIES \bullet empty_type_list(\mathrm{type}) = type?$

$inv_list' = inv_list \cup \{object?\}$

$inv_info' = inv_info \cup \{object? \mapsto empty_type_list\}$

Where *empty_type_list* is a predefined initial property list for the type of object added to the inventory. This helps assure that all property lists for objects of the same type hold the same properties.

The operation *Del-Inv-Object* is used to remove an object from the inventory.

Del-Inv-Object

$\Delta INVENTORY$
$object? : OBJ_NAME$

$object? \in inv_list$

$inv_list' = inv_list - \{object?\}$
$inv_info' = \{object?\} \lhd inv_info$

Since the list of properties is fixed for each type of object the ability to add or delete a property is not needed. However, it will be necessary to update the value of a property. This can be done with the *Update-Inv-Obj-Prop* operation.

Update-Inv-Obj-Prop

$\Delta INVENTORY$
$object? : OBJ_NAME$
$p? : PROP_NAME$
$v? : VALUE$

$object? \in inv_list$

$inv_list' = inv_list$
$Update\text{-}property[\ inv_info(object?)/pl,\ inv_info'(object?)/pl'\ ]$

The operation *Look-Inv-Obj-Prop* provides the ability to examine the value of an object property.

Look-Inv-Obj-Prop

$\Xi INVENTORY$
$object? : OBJ_NAME$
$p? : PROP_NAME$
$v! : VALUE$

$object? \in inv_list$

$Get\text{-}property[\ inv_info(object?)/pl,\ inv_info(object?)/pl'\ ]$

The operation *List-Inv-Obj-Prop* returns a list of the properties and property values associated with the specified object.

List-Inv-Obj-Prop

$\equiv$ *INVENTORY*
object? : *OBJ_NAME*
list! : $\mathbb{P}\{$ *PROP_NAME* $\times$ *VALUE* $\}$

---

object? $\in$ *inv_list*

List-properties[*inv_info*(*object?*)/*pl*, *inv_info*(*object?*)/*pl'*]

The operation *List-Inv-Objects* returns a list of all the objects in the inventory.

List-Inv-Objects

$\equiv$ *INVENTORY*
list! : $\mathbb{F}$ *OBJ_NAME*

---

list! = *inv_list*

4.2.6 Tasks

A project consists of a collection of tasks that must be performed. Identifying the necessary tasks, staffing, supporting, and scheduling these tasks is the responsibility of project management. Projects are often organized around project plans and project phases. Here, these concepts are formalized as tasks.

Project tasks form a hierarchy. There are major tasks such as *reverse engineer* the source code, *document* the target system, and *plan* the project. Each major task organizes a collection of sub-tasks. For example, to reverse engineer the source code, sub-tasks for collecting the source code files, static analysis of source code, and recovering a design are required.

Each task has a goal. Each task specifies a procedure to be used to achieve the goal of that task. A task procedure may be a loose set of guidelines to direct a human or group of humans through the task or it may be a specific set of directions. A task procedure may specify sub-tasks that must be done to fulfill the purpose of the current task.

A task serves several purposes within a project. The project manager creates and defines project tasks. The project manager uses tasks to represent the project plan. Tasks are used in project tracking. To the project manager tasks are objects that can be manipulated. Thus tasks are viewed as static objects. Tasks are also dynamic objects. Project tracking requires that the real start and stop dates for a task be recorded. The time personnel spend on a task must be recorded. Project personnel use tasks as directives. A task specifies to personnel what work is to be done and how it can be achieved. This is an interesting notion. A task provides a control structure for a portion of the work done during a project. People "execute" a task or sometimes a task can be automated and executed as a program. Sometimes, portions of a task can be automated, creating tools used by humans to achieve the goal of a task.

There is a collection of information associated with any task. This information divides into several groups, each of which can be formalized. The three groups explained below are: planning, tracking, and personnel information.

Each task fits into the project plan and is associated with planning information. Each task has a scheduled start and stop date. Each task has an estimate of the total number of hours that will be worked by personnel on the task. Each task has a pre-condition. The pre-condition states the conditions that must be satisfied before the task can begin. This may specify that approval from the project manager must be given, or that certain other tasks must first be completed. Each task has a post-condition that states the goal of the task. The post-condition must be satisfied for the task to be considered complete.

To formalize this planning information several sets must be defined. The relation AFTER specifies an ordering among dates, namely which dates come later than others. Project time, as in the number of hours worked or the estimated number of hours for a task, can be represented by a number, but for clarity we use the set TIME.

$$AFTER \mathrel{\widehat{=}} DATE \leftrightarrow DATE$$
$$TIME \mathrel{\widehat{=}} \mathbb{N}$$

Planning information for a task includes a description about how the task work will be done. This includes scheduling and planning for sub-tasks that must be done for a task to fulfill its goal. The information about how a task can be performed is stored as a procedure. The task procedure (as formalized by PROCEDURE) specifies how the task is to be achieved. The degree of guidance given by the procedure will vary. Some tasks can be automated by tools. Other tasks have steps that require human judgement or human action. Reviews are a good example. The steps of a review can be specified, but the work must be done by humans. However, the sub-task of notifying review participants can be automated by using e-mail.

Identifying the tools required by a task is part of project planning. For many tasks there is a collection of tools that can be used to help achieve the task goal. For example, a documentation task will need tools for editing, browsing, and text formatting. By specifying the tools needed by a task, permission to use these tools can be granted to the people assigned to the task. The set of all possible tools is denoted by TOOL.

$$TOOL \mathrel{\widehat{=}} \textit{The set of all possible tools}$$

Each tool has a name that uniquely identifies it. The set of all possible tool names is denoted by TOOL_NAME.

$$\begin{array}{|l} TOOL_NAME \mathrel{\widehat{=}} \textit{The set of all possible tool names} \\ \hline TOOL_NAME \subset NAME \end{array}$$

Task planning information is formalized by the schema TASK_PLAN.

$$
\begin{array}{l}
\textit{TASK_PLAN} \\
\hline
\textit{planned_start} : \textit{DATE} \\
\textit{planned_stop} : \textit{DATE} \\
\textit{planned_time} : \textit{TIME} \\
\textit{start_condition} : \textit{CONDITION} \\
\textit{goal} : \textit{CONDITION} \\
\textit{task_proc_name} : \textit{PROC_NAME} \\
\textit{tools} : \mathbb{F}\ \textit{TOOL_NAME} \\
\hline
\textit{planned_stop} \neq \epsilon \Rightarrow \textit{planned_stop}\ \textit{AFTER}\ \textit{planned_start} \\
\hline
\end{array}
$$

As a task is performed, project tracking will collect information about the task. The actual start and stop dates must be recorded. These dates may differ from the planned start and stop dates. The status of the task must be maintained. A task may be **inactive**, i.e. waiting for work to begin, **active** meaning that work is being done, or *completed* meaning that the task goal has been satisfied and no work on the task continues. In rare cases a task can be ***suspended*** due to a delay in a project. Task status information is denoted by the set TSTATUS.

$$\textit{TSTATUS} \mathrel{\widehat{=}} \{\ \textit{inactive},\ \textit{in_progress},\ \textit{completed},\ \textit{suspended}\ \}$$

Task tracking information is formalized by the schema TASK_TRACK.

$$
\begin{array}{l}
\textit{TASK_TRACK} \\
\hline
\textit{actual_start} : \textit{DATE} \\
\textit{actual_stop} : \textit{DATE} \\
\textit{status} : \textit{TSTATUS} \\
\hline
\textit{actual_stop} \neq \epsilon \Rightarrow \textit{actual_stop}\ \textit{AFTER}\ \textit{actual_start} \\
\textit{actual_stop} \neq \epsilon \Leftrightarrow \textit{status} = \textit{completed} \\
\textit{status} \neq \textit{inactive} \Leftrightarrow \textit{actual_start} \neq \epsilon \\
\hline
\end{array}
$$

During a project, personnel will be assigned to tasks. Assigning people to tasks occurs during project planning and throughout a project. Each person is identified by a name. The set of all people names is denoted by PERSON_NAME.

$$
\begin{array}{l}
\textit{PERSON_NAME} \mathrel{\widehat{=}} \textit{The set of all possible people names} \\
\hline
\textit{PERSON_NAME} \subset \textit{NAME}
\end{array}
$$

Each task has a person who is assigned as the task leader. Additional people may also be assigned to a task. Task planning includes an estimate of the amount of time each person will work on a task. Tracking information is also associated with personnel information. People spend time working on a task and the hours worked must be recorded. Personnel assigned to a task have a status. At first, personnel may be tentatively assigned, later personnel may become active on a task, and even later may finish the task, or be re-assigned to another project. The set of possible personnel statuses is denoted by PSTATUS.

$$\textit{PSTATUS} \mathrel{\widehat{=}} \{\ \textit{assigned},\ \textit{active},\ \textit{inactive},\ \ldots\ \}$$

For each person both planning and tracking information must be recorded. This information is formalized by the partial function PERSON_INFO. The tuple lists a person's status, estimated number of hours worked, and recorded number of hours worked.

$$PERSON_INFO \mathrel{\widehat{=}} PERSON_NAME \rightarrowtail PSTATUS \times TIME \times TIME$$

Information about personnel assigned to a task is formalized by the schema TASK_PERSONNEL. The predicate for this schema states that if the task leader has been assigned then information about the task leader must be present in the worker information function.

```
┌─ TASK_PERSONNEL ──────────────────────────────────────
│ task_leader : PERSON_NAME
│ workers : PERSON_INFO
├──────────
│ task_leader ≠ ε ⇒ task_leader ∈ dom workers
└───────────────────────────────────────────────────────
```

The parts described above are assembled into a schema that defines the properties of a task.

```
┌─ TASK ────────────────────────────────────────────────
│ TASK_PLAN
│ TASK_TRACE
│ TASK_PERSONNEL
├──────────
│ status = completed ⇒ goal
│ status ≠ inactive ⇒ task_leader ≠ ε
└───────────────────────────────────────────────────────
```

The first predicate states that the goal of a task must be satisfied before the task can be considered completed. The second predicate states that if the status has been started then a task leader must have been assigned.

Operations on PERSON_INFO

Before defining operations on tasks a few operations on task personnel information must be defined. Information about workers is modeled by a partial function that given an employee's name returns a tuple of information. Operations for extracting information from such a tuple are :

- Pstatus
- Est_time
- Rec_time

The *Pstatus* operation takes a personnel information tuple and returns the specified person's status.

Pstatus

$(s : PSTATUS,\ et : TIME,\ wt : TIME)?$

$status! : PSTATUS$

$status! = s$

The *Est_time* operation takes a personnel information tuple and returns the estimated work hours for the specified person.

Est_time

$(s : PSTATUS,\ et : TIME,\ wt : TIME)?$

$time! : TIME$

$time! = et$

The *Rec_time* operation takes a personnel information tuple and returns the recorded number of work hours for the specified person.

Rec_time

$(s : PSTATUS,\ et : TIME,\ wt : TIME)?$

$time! : TIME$

$time! = wt$

Task Operations

There are a number of necessary and useful operations that can be performed on a task. Operations on task planning information include:

- Create-task
- Assign-task-dates
- Assign-est-time
- Assign-start-condition
- Assign-task-goal
- Assign-task-procedure
- Get-task-dates
- Get-task-est-time
- Get-task-start-cond
- Get-task-goal

- Get-task-procedure
- Add-task-tool
- Delete-task-tool
- List-task-tools

Operations on task tracking information include:

- Start-task
- Stop-task
- Suspend-task
- Get-task-status

Operations on task personnel information include:

- Assign-task-leader
- Add-task-personnel
- Remove-task-personnel
- Set-task-personnel-status
- Get-task-personnel-info
- Record-personnel-time
- Get-total-task-time
- List-task-personnel

To denote that an operation on a TASK changes information about the TASK the following notation is used to denote which section of information is changed.

$$\Delta TASK_PLAN \mathrel{\widehat{=}} [\ TASK_PLAN, TASK_PLAN'\ ]$$

$$\Delta TASK_TRACK \mathrel{\widehat{=}} [\ TASK_TRACK, TASK_TRACK'\ ]$$

$$\Delta TASK_PERSONNEL \mathrel{\widehat{=}} [\ TASK_PERSONNEL, TASK_PERSONNEL'\ ]$$

To denote that an operation on a TASK does not change the TASK information the following notation is used:

$$\Xi TASK_PLAN \mathrel{\widehat{=}} [\ TASK_PLAN, TASK_PLAN'\ \mid\ TASK_PLAN' = TASK_PLAN\ ]$$

$$\equiv TASK_TRACK \mathrel{\widehat{=}} [\, TASK_TRACK, TASK_TRACK' \mid TASK_TRACK' = TASK_TRACK \,]$$

$$\equiv TASK_PERSONNEL \mathrel{\widehat{=}} [\, TASK_PERSONNEL, TASK_PERSONNEL' \mid TASK_PERSONNEL = TASK_PERSONNEL' \,]$$

The operation *Create-task* is used to create a new task. All task variables are set to specify that they are initially unassigned.

Create-task

$TASK$
$\Delta TASK_PLAN$
$\Delta TASK_TRACK$
$\Delta TASK_PERSONNEL$

$planned_start' = \epsilon$
$planned_stop' = \epsilon$
$planned_time' = 0$
$start_condition' = false$
$goal' = false$
$task_proc_name' = \epsilon$
$tools' = \varnothing$

$actual_start' = \epsilon$
$actual_stop' = \epsilon$
$status' = inactive$

$task_leader' = \epsilon$
$workers' = \varnothing$

Each element in the TASK schema needs to be set during a project. Thus an "assign value" operation is needed for each element. Only a few of these operations are given here since they are all similar. The first operation of this type is *Assign-task-dates*. This operation is used when project planning has determined the estimated starting and stopping times for a task.

Assign-task-dates

$TASK$
$\Delta TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$start? : DATE$
$stop? : DATE$

$planned_start' = start?$
$planned_stop' = stop?$

$planned_time' = planned_time$
$start_condition' = start_condition$
$goal' = goal$
$task_proc_name' = task_proc_name$
$tools' = tools$

The operation *Add-task-tool* adds a new tool to the list of tools that can by used by a task.

Add-task-tool

$TASK$
$\Delta TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$tool? : TOOL_NAME$

$tools' = tools \cup \{tool?\}$

$planned_start' = planned_start$
$planned_stop' = planned_stop$
$planned_time' = planned_time$
$start_condition' = start_condition$
$goal' = goal$
$task_proc_name' = task_proc_name$

The next "assign value" operation is *Start-task*. This operation records the date on which the task work started. It also verifies that the starting condition is satisfied. If everything is all right the task status is upgraded to **active**. Error handling is not currently specified, i.e. what if the start_condition is not satisfied.

Start-task

$TASK$
$\equiv TASK_PLAN$
$\Delta TASK_TRACK$
$\equiv TASK_PERSONNEL$
$start? : DATE$

$start_condition \Rightarrow actual_start' = start? \wedge status' = in_progress$
$\neg\, start_condition \Rightarrow actual_start' = actual_start \wedge status' = status$

$actual_stop' = actual_stop$

The operation *Assign-task-leader* is used to specify the task leader. The person who is to be the task leader must already be entered in the list of workers assigned to the task.

Assign-task-leader

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\Delta TASK_PERSONNEL$
$leader? : PERSON_NAME$

$leader? \in \mathrm{dom}\ workers \Rightarrow task_leader' = leader?$
$leader? \notin \mathrm{dom}\ workers \Rightarrow task_leader' = task_leader$

$workers' = workers$

The final example of an "assign value" operation is *Add-task-personnel*. This operation is used to assign a new person to a task. This operation can not be used to change information about a person already assigned to the task.

Add-task-personnel

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\Delta TASK_PERSONNEL$
$person? : PERSON_NAME$
$status? : PSTATUS$
$hours? : ASSIGNED_HOURS$

$person? \notin \mathrm{dom}\ workers \Rightarrow$
$\quad workers' = workers \cup \{ person? \mapsto (status?, hours?, 0) \}$
$person? \in \mathrm{dom}\ workers \Rightarrow workers' = workers$

$task_leader' = task_leader$

The operation *Stop-task* records the date on which the task work ends. A task can end when its goal has been satisfied.

Stop-task

$TASK \equiv TASK_PLAN$
$\Delta TASK_TRACK$
$\equiv TASK_PERSONNEL$
$stop? : DATE$

$goal \Rightarrow actual_stop' = stop? \ \wedge \ status' = completed$
$\not{g}oal \Rightarrow actual_stop' = actual_stop \ \wedge \ status' = status$

$actual_start' = actual_start$

The operation *Remove-task-personnel* is used to remove a person from a task. When a person is taken off a task, any hours worked by that person on that task must be saved, so only the status of the person is changed.

Remove-task-personnel

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\Delta TASK_PERSONNEL$
$person? : PERSON_NAME$

$person? \in \mathrm{dom}\ workers$
$workers' = workers \oplus \{ person? \mapsto$
$\quad (inactive,\ Est_time(workers(person?)),\ Rec_time(workers(person?)) \ )$

$task_leader' = task_leader$

The operation *Record-personnel-time* is used to log the hours worked by a person on a task.

Record-personnel-time

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\Delta TASK_PERSONNEL$
$person? : PERSON_NAME$
$newhours? : TIME$

$workers' = workers \oplus \{ person? \mapsto$
$\quad (Pstatus(workers(person?)),\ Est_hours(workers(person?)),$
$\quad\quad Rec_time(workers(person?)) + newhours? \ )$

$task_leader' = task_leader$

The operation *Get-task-procedure* is used to retrieve information about the procedure associated with a task.

Get-task-procedure

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$\equiv PROJECT - PROCEDURES$
$task_procedure! : PROC_NAME$

$task_procedure! = task_proc_name$

The operation *Get-task-status* is use to lookup the status of a task.

Get-task-status

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$state! : TSTATUS$

$state! = status$

The operation *Get-total-task-time* is used to calculate the total number of hours that have been worked on a task. This number is determined by summing the hours worked by each person who has worked on the task.

Get-total-task-time

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$total_time! : TIME$

$total_time! = \sum_{w \in \mathrm{dom}\, workers} Rec_time(\ workers(w)\ )$

The operation *List-task-personnel* is used to show all personnel assigned to a task.

List-task-personnel

$TASK$
$\equiv TASK_PLAN$
$\equiv TASK_TRACK$
$\equiv TASK_PERSONNEL$
$personnel! : \mathbb{F}\ PERSON_NAME$

$personnel! = \mathrm{dom}\ workers$

4.2.7 Project Plan

The project plan defines the project structure. The plan specifies how work will be conducted, determines phases of the project, tasks involved in each phase, and how the tasks are to be done. The plan states how the target system will be produced. Defining the project plan is a major portion of the project planning work.

The basic unit of project planning is the task. The project approach defines the tasks that must be done to reach the target system. The concept of a task has been formalized by the schema TASK. Each task is given a name. The set TASK_NAME denotes the set all possible task names.

$$\begin{array}{|l} TASK_NAME \mathrel{\widehat{=}} \textit{The set of all possible task names} \\ \hline TASK_NAME \subset NAME \end{array}$$

The set of all tasks to be used within a project is denoted by PROJECT-TASK.

$$PROJECT\text{-}TASK : TASK_NAME \rightarrowtail\!\!\!\!\rightarrow TASK$$

Notice that is a one-to-one partial function. This states that each task in a project is unique. Some tasks may occur several times in a project, such as document-program, but these tasks will be performed on different data each time, making the different instances of a task unique.

Initially, there are no tasks planned for a project. The initial collection of project tasks is denoted by INIT-PROJECT-TASK

$$INIT\text{-}PROJECT\text{-}TASK = \varnothing$$

A large project is typically divided into phases. For re-engineering typical phases include reverse engineering, re-design, re-implementation, testing, system transition, and documentation. In addition, a large system may be re-engineered in pieces, with one portion of the system re-engineered before work begins on the next. An entire system may be re-engineered in increments, where each increment produces a better version of the system and slowly converges toward the desired target system.

A phase is considered to be a set of tasks. All the tasks in a phase are presumed to work towards a common phase goal. The concept of a project phase is denoted by PHASE.

$$\begin{array}{|l} PHASE \subseteq \mathrm{dom}\, PROJECT\text{-}TASK \\ \hline PHASE \neq \varnothing \end{array}$$

The project plan is modeled by a partial function from names to phases. This function is used to name each phase. In addition, the name assigned to a phase matches the name of a task. The task with this name is the phase task and the directions for performing the phase are contained within the task information.

$$\begin{array}{|l} \underline{PROJECT\text{-}PLAN} \\ plan : TASK_NAME \rightarrow PHASE \\ \hline \mathrm{dom}\, plan \subseteq \mathrm{dom}\, PROJECT\text{-}TASK \end{array}$$

When the project starts the PROJECT-PLAN is already initialized. The plans specifies one phase, the analysis phase, and only those tasks that are used in that phase. One result of defining the project approach is that the project plan is expanded to specify the entire project. The issue of how the initial project plan is defined is left as outside the scope of the re-engineering process.

Project Plan Operations

There are several operations on the project plan that can be defined. These are:

- Create-plan-phase
- Delete-plan-phase
- Get-plan-phase
- List-plan-phases
- Add-phase-task
- Delete-phase-task
- List-phase-tasks

There are several operations that can be define on the project tasks also. These are:

- Create-proj-task
- Delete-proj-task
- Get-proj-task
- List-proj-tasks

To denote that an operation changes information about the project tasks or the project plan the following notation is used:

$$\Delta PROJECT\text{-}TASK \mathrel{\hat{=}} [\, PROJECT\text{-}TASK, PROJECT\text{-}TASK' \,]$$

$$\Delta PROJECT\text{-}PLAN \mathrel{\hat{=}} [\, PROJECT\text{-}PLAN, PROJECT\text{-}PLAN' \,]$$

To denote that an operation does not change information about the project tasks or the project plan the following notation is used:

$$\Xi PROJECT\text{-}TASK \mathrel{\hat{=}} [\, PROJECT\text{-}TASK, PROJECT\text{-}TASK' \mid PROJECT\text{-}TASK' = PROJECT\text{-}TASK \,]$$

$$\Xi PROJECT\text{-}PLAN \mathrel{\hat{=}} [\, PROJECT\text{-}PLAN, PROJECT\text{-}PLAN' \mid PROJECT\text{-}PLAN = PROJECT\text{-}PLAN \,]$$

The operation *Create-plan-phase* creates a new phase and adds it to the project plan. Remember each phase is a collection of task names and each phase is named after a major task. The task that a phase is named after is part of the phase.

Create-plan-phase

$\Delta PROJECT\text{-}PLAN$
$\Xi PROJECT\text{-}TASK$
$task? : TASK_NAME$

$task? \in \mathrm{dom}\, PROJECT\text{-}TASK$
$task? \notin \mathrm{dom}\, plan$
$plan' = plan \cup \{ task? \mapsto \{task?\}\}$

The operation *Delete-plan-phase* deletes a phase from the project plan.

Delete-plan-phase

$\Delta PROJECT\text{-}PLAN$
$phase? : TASK_NAME$

$phase? \in \mathrm{dom}\, plan$
$plan' = \{phase?\} \lhd plan$

The operation *Add-phase-task* adds a task to a phase.

Add-phase-task

$\Delta PROJECT\text{-}PLAN$
$\Xi PROJECT\text{-}TASK$
$phase? : TASK_NAME$
$task? : TASK_NAME$

$task? \in \mathrm{dom}\, PROJECT\text{-}TASK$
$phase? \in \mathrm{dom}\, plan$
$task? \notin plan(phase?)$
$plan' = plan \oplus \{ phase? \mapsto plan(phase?) \cup \{task?\} \}$

The operation *Create-proj-task* creates a new task and associates it with the list of project tasks.

Create-proj-task

$\Delta PROJECT\text{-}TASK$
$task? : TASK_NAME$

$task? \notin \mathrm{dom}\, PROJECT\text{-}TASK$
$PROJECT\text{-}TASK' = PROJECT\text{-}TASK \cup \{ task? \mapsto Create\text{-}task \}$

The operation *Delete-proj-task* removes a task from the list of project tasks. A task can not be

$$
\begin{array}{l}
\textit{Delete-proj-task} \\
\hline
\Delta PROJECT\text{-}TASK \\
\Xi PROJECT\text{-}PLAN \\
task? : TASK_NAME \\
\hline
task? \in \mathrm{dom}\, PROJECT\text{-}TASK \\
task? \notin \mathrm{dom}\, PROJECT\text{-}PLAN \\
\forall p : TASK_NAME \bullet p \in \mathrm{dom}\, PROJECT\text{-}PLAN \wedge task? \notin PROJECT\text{-}PLAN(p) \\
PROJECT\text{-}TASK' = \{task?\} \lhd PROJECT\text{-}TASK \\
\hline
\end{array}
$$

4.2.8 Acceptance Criteria

The project team and the client must reach an agreement that specifies the conditions under which the target system will be accepted. These conditions are often called the acceptance criteria. It is not sufficient to state that the target system will successfully pass all system tests or that the target system will correctly implement the system requirements. The acceptance criteria should complement, not duplicate the internal quality assurance procedures. Acceptance criteria can specify a broad range of conditions that must be satisfied before the target system will be accepted. Possible criteria include:

- Functionality of the existing system that was not scheduled to be changed (i.e, the requirements for the feature are the same in both the existing and the target specifications) will operate identically in both the existing and target systems.
- Each document for the existing system will either be rewritten completely or updated to account for differences between the existing and target systems.
- The test cases used will execute a specified percentage of the system source code.
- Capability xyz of the target system will execute within its timing constraints for the following test cases ...
- The target system will have a reliability level such that it only fails once per specified unit of time.
- Changes to the user interface will be documented.

The set of all possible criteria is denoted as CRITERIA.

$$CRITERIA \mathrel{\widehat{=}} \textit{The set of all possible criteria}$$

The concept of acceptance criteria is formalized as an one-to-one partial function from labels to criteria. Each criterion in the acceptance criteria must have a unique label. The nature of a criterion is vague. A criterion is a requirement or condition that must be satisfied by the system and related products, such as documentation.

$$ACCEPT_CRITERIA : LABEL \rightarrowtail CRITERIA$$

Initially, no acceptance criteria is specified. The initial acceptance criteria is denoted by INIT_ACCEPT_CRIT.

$$INIT_ACCEPT_CRITERIA = \varnothing$$

Acceptance Criteria Operations

There are several basic operations that can operate on the acceptance criteria information. These operations are:

- Add-accept-criterion
- Delete-accept-criterion
- Update-accept-crit-label
- Update-accept-criterion
- Get-accept-criterion
- List-accept-criteria

To denote that an operation changes the acceptance criteria information the Δ notation is used:

$$\Delta ACCEPT_CRITERIA \mathrel{\widehat{=}} [\, ACCEPT_CRITERIA,\ ACCEPT_CRITERIA' \,]$$

To denote that an operation does not change the acceptance criteria information the Ξ notation is used:

$$\Xi ACCEPT_CRITERIA \mathrel{\widehat{=}} [\, ACCEPT_CRITERIA,\ ACCEPT_CRITERIA' \mid ACCEPT_CRITERIA' = ACCEPT_CRITERIA \,]$$

The operation *Add-accept-criterion* is used to add a new criterion to the list of acceptance criteria.

Add-accept-criterion

$$\begin{array}{l}
\Delta ACCEPT_CRITERIA \\
label? : LABEL \\
crit? : CRITERIA \\
\hline
label? \notin \mathrm{dom}\ ACCEPT_CRITERIA \\
crit? \notin \mathrm{ran}\ ACCEPT_CRITERIA \\
ACCEPT_CRITERIA' = ACCEPT_CRITERIA \cup \{\, label? \mapsto crit? \,\}
\end{array}$$

The operation *Delete-accept-criterion* is used to remove a criterion from the listed acceptance criteria.

Delete-accept-criterion

$\Delta ACCEPT_CRITERIA$
$label? : LABEL$

$label? \in \mathrm{dom}\ ACCEPT_CRITERIA$
$ACCEPT_CRITERIA' = \{ label? \} \lhd ACCEPT_CRITERIA$

The operation *Update-accept-crit-label* is used to change the label assigned to a criterion.

Update-accept-crit-label

$\Delta ACCEPT_CRITERIA$
$old_label : LABEL$
$new_label? : LABEL$

$old_label? \in \mathrm{dom}\ ACCEPT_CRITERIA$
$new_label? \notin \mathrm{dom}\ ACCEPT_CRITERIA$
$ACCEPT_CRITERIA' = \{ old_label? \} \lhd ACCEPT_CRITERIA \cup$
$\{ new_label? \mapsto ACCEPT_CRITERIA(old_label) \}$

The operation *Update-accept-criterion* is used to change the criterion associated with a label. The idea is that a criterion has been updated, but not completely discarded and replaced.

Update-accept-criterion

$\Delta ACCEPT_CRITERIA$
$label? : LABEL$
$crit? : CRITERIA$

$label? \in \mathrm{dom}\ ACCEPT_CRITERIA$
$crit? \notin \mathrm{ran}\ ACCEPT_CRITERIA$
$ACCEPT_CRITERIA' = ACCEPT_CRITERIA \oplus \{ label? \mapsto crit? \}$

The operation *Get-accept-criterion* is used to retrieve the criterion associated with the specified label.

Get-accept-criterion

$\Xi ACCEPT_CRITERIA$
$label? : LABEL$
$crit! : CRITERIA$

$label? \in \mathrm{dom}\ ACCEPT_CRITERIA$
$crit! = ACCEPT_CRITERIA(label?)$

The operation *List-accept-criteria* is used to obtain a list of all the acceptance criteria.

List-accept-criteria

$\equiv ACCEPT_CRITERIA$

$list! : \mathbb{F}\{ LABEL \times CRITERIA \}$

$list! = \{ l : label;\ c : CRITERIA \mid ACCEPT_CRITERIA(l) = c \}$

4.2.9 Project Procedures

One result of project planning is the development and/or collection of procedures that are to be followed during a project. Many software organizations typically have a set of standard procedures that define a way of doing things within the organization. Procedures may describe the correct way to extract information from a configuration management system, provide directions about how to report errors, or describe how to set up and conduct a review session.

Many re-engineering tasks are highly proceduralized and require little human judgement. However, procedures may need fine-tuning to accommodate project-specific details. The specification and use of project procedures helps to ensure consistency during the project. This can simplify the burden of managing a project and is reflected in the consistency and overall quality of the target system. Some typical procedures are listed below:

- Inventory collection
- Testing
- Project monitoring
- Project reviews
- Reverse engineering of source code
- Reverse engineering of data files
- Installing target system
- Change control
- Error reporting
- Quality assurance
- Status reporting

Procedures are used throughout any project. An ideal procedure provides detailed step-by-step instructions for performing a task. Such a procedure can be automated and incorporated into a tool or operation used to support a project. Unfortunately, many procedures used during a project can only be loosely stated and require either human guidance or must be done manually.

The general characteristics of procedures can be formally specified. A procedure is a sequence of steps. Typically, there is an ordering among steps, where certain steps can not be started until

other steps are completed. A procedure always has a step designated as the first step. The first step is the step that is initiated when the procedure begins. For convenience it is assumed that steps are labeled. Further, a step consists of a label, a pre-condition, an action, and a post-condition. The concept of actions is formalized by the set ACTION. For now the exact details of what constitutes an action will be left unspecified.

$$ACTION \mathrel{\widehat{=}} \textit{The set of all possible actions}$$

The concept of a procedure step is formalized by the set STEP. The definition of STEP says that each step has an identifying label. Each step has two conditions. The first condition is a pre-condition, a condition that must hold true before the action associated with the step can be performed. The second condition is a post-condition, a condition that must hold true after the step action is complete.

$$STEP \mathrel{\widehat{=}} LABEL \times CONDITION \times ACTION \times CONDITION$$

The concept of a procedure is formalized by the schema PROCEDURE. A procedure consists of a finite number of steps, an ordering among the steps, and a designated first step.

$$\begin{array}{l}
\textit{PROCEDURE} \\
\hline
proc : \mathbb{F}\, STEP \\
method : LABEL \rightarrow LABEL \\
first\text{-}step : LABEL \\
\hline
\forall\, l_1, l_2 : LABEL,\ a_1, a_2 : ACTION,\ c_1, c_2, p_1, p_2 : CONDITION \bullet \\
\quad l_1\ method\ l_2 \Rightarrow (\ \exists (l_1, c_1, a_1, p_1) \in proc \wedge \exists (l_2, c_2, a_2, p_2) \in proc\ ) \\
\forall\, l : LABEL,\ a : ACTION,\ c_1, c_2 : CONDITION \bullet \\
\quad (l, c_1, a, c_2) \in proc \Rightarrow (\exists\, l_1 : LABEL \mid l_1\ method\ l \vee l\ method\ l_1) \\
\exists\, a : ACTION,\ c_1, c_2 : CONDITION \bullet (first\text{-}step, c_1, a, c_2) \in proc \\
\hline
\end{array}$$

The variable **method** is a relation that specifies the relationship between steps in the procedure. This relationship gives an ordering between steps that specifies those steps that can be done after any given step.

Because it is not always possible to specify a procedure for every task, we define a special procedure called the *NULL-PROCEDURE*.

$$\begin{array}{l}
\textit{NULL-PROCEDURE} \\
\hline
PROCEDURE \\
\hline
proc = \varnothing \\
method = \varnothing \\
first\text{-}step = \epsilon \\
\hline
\end{array}$$

When collecting procedures to be used in a project it is convenient to name each procedure. Each procedure should be well documented. Procedure documentation should explain the purpose

of a procedure, how it is used, considerations, possible problems that may be encountered and how to solve them, etc. At this point, we will define the procedure document as a piece of text and leave the details of structuring the document for later.

When a procedure is collected or developed for a project the procedure must be named, documented, and the procedure itself specified. A procedure with these three components is called a complete procedure. The set of all possible procedure names is denoted by PROC_NAME.

$$\begin{array}{|l} PROC_NAME \mathrel{\widehat{=}} \textit{The set of all possible procedure names} \\ \hline PROC_NAME \subset NAME \end{array}$$

The set of all possible complete procedures is called COMPLETE-PROCEDURE.

$$COMPLETE\text{-}PROCEDURE \mathrel{\widehat{=}} PROC_NAME \times TEXT \times PROCEDURE$$

The collection of all complete procedures to be used during a project is called PROJECT-PROCEDURES and is denoted as:

$$\begin{array}{|l} PROJECT\text{-}PROCEDURES : \mathbb{F}\, COMPLETE\text{-}PROCEDURES \\ \hline (\epsilon, \varnothing, NULL\text{-}PROCEDURE) \in PROJECT\text{-}PROCEDURES \end{array}$$

Initially, no procedures are associated with a project. The initial collection of procedures is denoted by INIT-PROJ-PROCS.

$$INIT\text{-}PROJ\text{-}PROCS = (\epsilon, \varnothing, NULL\text{-}PROCEDURE)$$

Procedure Operations

Since project procedures are merely a set of complete procedures the only operations that are necessary are :

- Delete-proj-procedure
- Get-proj-procedure
- Update-proj-procedure
- List-proj-procs

Several operations on complete procedures can also be defined. These are:

- Create-complete-proc
- Update-cproc-name
- Update-cproc-text

- Update-cprocedure
- Get-cproc-name
- Get-cproc-doc
- Get-cprocedure

To denote that an operation changes information about a procedure the Δ notation is used:

$$\Delta PROJECT\text{-}PROCEDURES \mathrel{\widehat{=}} [\, PROJECT\text{-}PROCEDURES, \; PROJECT\text{-}PROCEDURES \,]$$

To denote that an operation does not change information about a procedure the $\equiv$ notation is used:

$$\equiv PROJECT\text{-}PROCEDURES \mathrel{\widehat{=}} [\, PROJECT\text{-}PROCEDURES, \; PROJECT\text{-}PROCEDURES \mid PROJECT\text{-}PROCEDURES' = PROJECT\text{-}PROCEDURES \,]$$

The operation *Delete-proj-procedure* deletes a complete procedure from the list of project procedures.

Delete-proj-procedure

$$\Delta PROJECT\text{-}PROCEDURES$$
$$proc? : PROC_NAME$$

$$\exists\, t : TEXT;\; p : PROCEDURE \bullet (proc?, t, p) \in PROJECT\text{-}PROCEDURES$$
$$PROJECT\text{-}PROCEDURES' = PROJECT\text{-}PROCEDURES - \{(proc?, t, p)\}$$

The operation *List-proj-procs* returns a list of procedure names that have been specified for a project.

List-proj-procs

$$\equiv PROJECT\text{-}PROCEDURES$$
$$names! : \mathbb{F}\, PROC_NAME$$

$$names! = \{n : PROC_NAME \mid \exists\, t : TEXT;\; p : PROCEDURE \bullet (n, t, p) \in PROJECT\text{-}PROCEDURES\}$$

The operation *Create-complete-proc* is used to create a new complete procedure and enter it into the list of project procedures.

Create-complete-proc

$$\Delta PROJECT\text{-}PROCEDURES$$
$$name? : PROC_NAME$$

$$\neg\, \exists\, t : TEXT;\; p : PROCEDURE \bullet (name?, t, p) \in PROJECT\text{-}PROCEDURES$$
$$PROJECT\text{-}PROCEDURES' = PROJECT\text{-}PROCEDURES \cup \{(name?, \epsilon, NULL\text{-}PROCEDURE)\}$$

The operation *Update-cprocedure* changes the procedure associated with a complete procedure name.

Update-cprocedure

$\Delta PROJECT\text{-}PROCEDURE$
$name? : PROC_NAME$
$proc? : PROCEDURE$

$\exists\, t : TEXT;\ p : PROCEDURE \bullet (name?, t, p) \in PROJECT\text{-}PROCEDURES$
$PROJECT\text{-}PROCEDURES' = (\ PROJECT\text{-}PROCEDURES - (name?, t, p)\ ) \cup \{(name?, t, proc?)\}$

4.2.10 Project Standards

One result of project planning is the collection and/or definition of standards that are to be applied to project products. Many software organizations typically have a set of standards that are adhered to within the organization. In addition, new standards may be needed for a new project and some existing standards may need to be tailored to satisfy project specific requirements.

Standards define the accepted form of results for each task in the process. Standards place constraints on what is and is not acceptable. Typical project standards include :

- Naming standards (variables, functions, files)
- Documentation standards
- Coding standards
- Test standards

Standards tend to be lists of rules that identify some item or property of an object and then place constraints on that item. Typical constraints limit the range of valid values or the forms that may be used to represent the item. We formalize the concepts of items and constraints by denoting them with the sets:

$ITEM \mathrel{\widehat{=}}$ *The set of all possible items*
$CONSTRAINT \mathrel{\widehat{=}}$ *The set of all constraints*

The set ITEM denotes the set of all possible items. The set CONSTRAINT denotes the set of all possible constraints. For one item there may be several constraints, this is denoted by a constraint list. A constraint list is a set of constraints.

$CONSTRAINT\text{-}LIST : \mathbb{P}\, CONSTRAINT$

A rule identifies an item and specifies the constraints on that item. For convenience each rule is given a name. The set of all possible rule names is RULE_NAME.

$RULE_NAME \mathrel{\widehat{=}}$ *The set of all possible rule names*

$RULE_NAME \subset NAME$

$RULE : RULE_NAME \times ITEM \times CONSTRAINT\text{-}LIST$

A standard is defined as a set of rules.

$$
\begin{array}{|l}
\hline
STANDARD \\
standard : \mathbb{F}\, RULE \\
\hline
\forall n_1 : RULE_NAME,\ i_1, i_2 : ITEM,\ c_1, c_2 : CONSTRAINT\text{-}LIST \bullet \\
\quad (n_1, i_1, c_1) \in standard \Rightarrow \neg\ (\ \exists (n_1, i_2, c_2) \in standard\ \wedge\ i_1 \neq i_1\ \wedge\ c_1 \neq c_2\ ) \\
\hline
\end{array}
$$

The collection of project standards is given by PROJECT-STANDARDS.

$PROJECT\text{-}STANDARDS : \mathbb{F}\, STANDARD$

Standard Operations

Since the project standards is merely a set of standards the only operations that are necessary are:

- Delete-proj-standard
- Get-proj-standard
- Update-proj-standard
- List-proj-standards

Several operations on standards can also be defined. These are:

- Create-standard
- Update-stand-name
- Update-stand-item
- Add-stand-constraint
- Delete-stand-constraint
- Get-rule
- List-rules

These operations are not be specified at this time.

4.2.11 Project Resources

The software system being re-engineered may execute on special equipment, or special equipment may be required to test the system, or even to aid in the re-engineering project. Perhaps a laser printer and several workstations are to be dedicated to the project. Any resource needed by a project must be identified and arrangements made to obtain the resource. Typical resources include:

- Computers
- Printers
- Peripheral devices
- Network access
- Simulation equipment - such as a mock-up cockpit
- Knowledge sources - key personnel not assigned to the project. This can include a system administrator, a developer of the existing system, a key system user, a person with hardware expertise, etc.

The list of resources is modeled by a partial function. Each resource has a name and an associated set of properties. Different resources have different properties.

$$RES_NAME \mathrel{\widehat{=}} \textit{The set of all possible resource names}$$
$$RES_NAME \subset NAME$$

The resource information is denoted by the partial function, RESOURCES, which maps the name of the resource to its list of properties.

$$RESOURCES : RES_NAME \rightarrowtail PROP_LIST$$

Initially, the resources required by a project are unknown. The initial project resources are formalized by *INIT-RESOURCES*.

$$INIT\text{-}RESOURCES = \varnothing$$

Resource Operations

Several operations can be defined on resource information. These operations include.

- Add-new-resource
- Delete-resource
- List-resources
- List-resource-props

- Add-resource-prop
- Delete-resource-prop
- Change-resource-prop

To denote that an operation changes the resource information the Δ notation is used:

$$\Delta RESOURCES \mathrel{\widehat{=}} [\, RESOURCES,\ RESOURCES' \,]$$

To denote that an operation does not change the resource information the Ξ notation is used:

$$\Xi\, RESOURCES \mathrel{\widehat{=}} [\, RESOURCES,\ RESOURCES' \mid RESOURCES' = RESOURCES \,]$$

The operation *Add-new-resource* is used to associate a resource with the resource information. The new resource should not already be listed.

Add-new-resource

$\Delta RESOURCES$
$name? : RES_NAME$

$name? \notin RESOURCES$
$RESOURCES' = RESOURCES \cup \{ name? \mapsto \varnothing \}$

The operation *Delete-resource* is used to remove a resource from the list of resources. The may be used when project emphasis shifts and previously needed equipment will no longer be used.

Delete-resource

$\Delta RESOURCES$
$name? : RES_NAME$

$name? \in \mathrm{dom}\ RESOURCES$
$RESOURCES' = \{name?\} \mathbin{\lhd} RESOURCES$

The operation *List-resources* is used to generate a list of all resources needed by the project.

List-resources

$\Xi\, RESOURCES$
$list! : \mathbb{P}\, RES_NAME$

$list! = \mathrm{dom}\ RESOURCES$

The operation *List-resource-props* is used to generate a list of properties associated with a resource item.

List-resource-props

$\equiv RESOURCES$
$name? : RES_NAME$
$list! : \mathbb{P}\{ PROP_NAME \times VALUE \}$

$name? \in \mathrm{dom}\ RESOURCES$
$List\text{-}properties[RESOURCES(name?)/pl,\ RESOURCES'(name?)/pl']$

The operation *Add-resource-prop* is used to add a new property to a listed resource.

Add-resource-prop

$\Delta RESOURCES$
$name? : RES_NAME$
$prop? : PROP_NAME$
$val? : VALUE$

$name? \in \mathrm{dom}\ RESOURCES$
$RESOURCES' = RESOURCES \oplus \{ name? \mapsto$
$Add\text{-}property[RESOURCES(name?)/pl, RESOURCES'(name?)/pl',$
$prop?/p?, val?/v?] \}$

The operation *Delete-resource-prop* is used to remove a property from a listed resource.

Delete-resource-prop

$\Delta RESOURCES$
$name? : RES_NAME$
$prop? : PROP_NAME$

$name? \in \mathrm{dom}\ RESOURCES$
$RESOURCES' = RESOURCES \oplus \{ name? \mapsto$
$Delete\text{-}property[RESOURCES(name?)/pl, RESOURCES'(name?)/pl',$
$prop?/p?] \}$

The operation *Change-resource-prop* is used to change the value associated with a property of a listed resource.

Change-resource-prop

$\Delta RESOURCES$
$name? : RES_NAME$
$prop? : PROP_NAME$
$val? : VALUE$

$RESOURCES' = RESOURCES \oplus \{ name? \mapsto$
$Update\text{-}property[RESOURCES(name?)/pl, RESOURCES'(name?)/pl',$
$prop?/p?, val?/v?] \}$

4.2.12 Project Tools

Tools can be used to implement tasks that are suitable for automation. Actions that are repetitive are candidates for automation by a tool. Tools are commonly used to record information and operate on that information. The use of tools can reduce project time, insure accuracy of results, and promote consistency.

During the planning phase tools may be identified that are needed on a project. Tools required for a project may already exist in-house or may need to be purchased, or as a last resort developed. The development of new tools for a project is often discouraged since the time required to develop a tool can delay a project. Typical project tools include:

- Language compilers
- Text and code editors
- Configuration management system
- Language translators
- Software engineering environment
- Documentation management system
- Documentation browsers
- Source code browsers
- File converters
- Test case generators
- Test coverage monitors
- File comparators
- Source code formatters
- Spreadsheets
- Schedule handlers
- Diagraming tools

For tools that will be used during a project, information about these tools must be collected. This information can be used during the project to guide the use of these tools and identify tasks that will require these tools.

Each tool has a name that identifies it. The set of all possible tool names is denoted by TOOL_NAME.

Each tool has a computer system that is resides on and a location on that system where it is stored. It is possible for a tool to exist on more than one system. Here, each system that has a

copy of the tool must be noted. Each tool has an available status. Typically, a tool is available full-time, but some unique tools may be available only on a limited or scheduled basis. A tool may also be on order, in development, or in-house but not installed. The availability of required tools can affect the project schedule. Each tool is used by one or task project tasks.

Before formalizing the concept of tool information it is necessary to define several sets. The set of all computers in an organization is denoted by COMPUTER.

$$COMPUTER \mathrel{\widehat{=}} \textit{The set of all possible organization computers}$$

The set of all possible directories or places to store a tool on these computers is denoted by BIN.

$$BIN \mathrel{\widehat{=}} \textit{The set of all tool locations}$$

The set of all possible availability statuses for a tool is denoted by TOOL_STATUS:

$$TOOL_STATUS \mathrel{\widehat{=}} \{ \textit{available-full}, \textit{available-scheduled}, \textit{on-order}, \textit{to-be-ordered}, \textit{not-installed} \}$$

The schema below formalizes the concept of tool information.

$$\begin{array}{|l}
\hline
\textit{TOOL} \\
\textit{name} : TOOL_NAME \\
\textit{location} : \mathbf{F}\ COMPUTER \times BIN \\
\textit{availability} : TOOL_STATUS \\
\textit{used_by} : \mathbf{F}\ TASK_NAME \\
\hline
\textit{used_by} \subseteq \mathrm{dom}\, PROJECT\text{-}TASK \\
\textit{used_by} = \{ t : TASK_NAME \mid t \in \mathrm{dom}\, PROJECT\text{-}TASK \wedge \\
\qquad \textit{name} \in PROJECT\text{-}TASK(t).tools \} \\
\hline
\end{array}$$

The collection of all tool information for a project is formalized by PROJECT-TOOLS, where :

$$PROJECT\text{-}TOOLS : \mathbf{P}\ TOOL$$

Initially, the tools to be used on a project are unknown. This is formalized by INIT-PROJECT-TOOLS.

$$INIT\text{-}PROJECT\text{-}TOOLS = \varnothing$$

Operations on Tool Information

The following operations can be defined on tool information and PROJECT-TOOLS.

- Add-tool-location
- Delete-tool-location

- Set-tool-availability
- Get-tool-name
- List-tool-location
- Get-tool-availability
- Get-tool-task-list

The following operations can be defined on the list of project tools.

- Add-project-tool
- Get-project-tool
- Update-project-tool
- Delete-project-tool
- List-project-tools

The specifications for these operations are not given.

4.2.13 Test Planning

The testing of programs and software systems is done to show that system features work, to search the system for faults, and to establish confidence that a system is reliable enough to be put into usage. Testing of a system occurs after the implementation phase when coding is complete. However, preparations for testing begin much earlier in a project. Test planning can begin during the early stages of a project.

The result of test planning is a test plan. A test plan specifies the tests that are used to verify that the system features operate without errors. Tests specified in a test plan can be developed using different strategies. Two common strategies are black-box and white-box testing. Black-box testing ignores source code and focuses on testing system features. White-box testing focuses on the source code. Two common white-box testing strategies are branch coverage and statement coverage. Branch coverage tries to execute all branches in the code. Statement coverage tries to execute all statements in the code. Typically in white-box testing a certain percentage of coverage is required. Regardless of the strategy or method used to create test cases, the basic properties of tests and test plans can be formalized.

A test case consists of input data and expected output results. The input data may be used to test a routine within a program, an entire program, or the interaction between programs (i.e. the system). The output results are the expected result to be produced by a test. The expected result can be compared against the actual results to determine if the test worked correctly. Input data may consist of keystrokes, output data may consist of recorded screens. Either input data or output results may consist of files, formatted text, the values assigned to specific source code variables,

signals sent or received to other programs or devices, etc. These types of various information forms are formalized by the set:

$$TEST_FORM \mathrel{\widehat{=}} \textit{The set of all possible input and output data}$$

A test case must also specify what is being tested by the test data. The item tested may the system, a program within the system, a routine with a program, a path through a program, a program statement, a feature of a program or the system, etc. The set of all system items that can be tested is denoted by:

$$TEST_ITEM \mathrel{\widehat{=}} \textit{The set of all testable items}$$

Finally, a test case should be documented. The test case documentation should explain the test, how the test is done, i.e using a test tool, manual entry, etc, and any considerations relevant to the test. Test case instructions are formalized as a COMPLETE-PROCEDURE. The set of all test cases is formalized by:

$$\begin{array}{l}
\textit{TEST_CASE} \\
\hline
test : LABEL \\
item : TEST_ITEM \\
input : TEST_FORM \\
result : TEST_FORM \\
proc : PROC_NAME \\
\hline
\end{array}$$

The label element is an unique label identifying the specific test case.

It is common to group tests together. This is often done to a set of tests that have some common thread, such as they test the same feature or routine. A test group should be named for easy reference. A test group should be documented, explaining the purpose the test grouping. Finally, there is often an ordering among test cases in a group, where the ordering may specify the order of test case execution. The group documentation and ordering information are formalized as part of a test group procedure using COMPLETE-PROCEDURE. The procedure will specify the order in which test cases are applied. The concept of a test group is formalized by the set TEST_GROUP.

$$\begin{array}{l}
\textit{TEST_GROUP} \\
\hline
group_name : LABEL \\
group_proc : PROC_NAME \\
tests : \mathbb{F}\ TEST_CASE \\
\hline
\neg\ \exists (l_1, i, in_1, out_1, p_1), (l_2, i_2, in_2, out_2, p_2) \in tests \bullet\ l_1 = l_2 \wedge \\
\qquad\qquad (i_1 \neq i_2 \vee in_1 \neq in_2 \vee out_1 \neq out_2 \vee p_1 \neq p_2)\ \} \\
\hline
\end{array}$$

Finally, test groups themselves are often grouped together to form a test plan. There is typically only one test plan for a system. The plan should be documented explaining the plan and the purpose of the tests. This should be high level test information. Details can be recorded in the group or test case documentation. Finally, an ordering between groups may be given, where the ordering specifies

which groups should be tested first. This ordering and test plan documentation are formalized by a test plan procedure using COMPLETE-PROCEDURE. The concept of a test plan is formalized by the set TEST_PLAN.

$$
\begin{array}{|l}
\hline
\textit{TEST_PLAN} \\
\textit{test_plan_proc} : \textit{COMPLETE - PROCEDURE} \\
\textit{groups} : \mathbb{F}\ \textit{TEST_GROUP} \\
\hline
\neg\ \exists (l_1, p_1, ts_1,),\ (l_2, p_2, ts_2) \in \textit{groups}\ \bullet \\
\qquad\qquad l_1 = l_2 \wedge (p_1 \neq p_2 \vee ts_1 \neq ts_2)\} \\
\hline
\end{array}
$$

Initially there is no test plan for a re-engineering project. There may be an existing test plan for the existing system, but such a plan must be evaluated before a decision is made to use it during the re-engineering project. The initial test plan for the target system is denoted by INIT_TEST_PLAN.

$$
\begin{array}{|l}
\hline
\textit{INIT_TEST_PLAN} \\
\textit{TEST_PLAN} \\
\hline
\textit{proc} = (\epsilon, \epsilon, \textit{NULL-PROCEDURE}) \\
\textit{groups} = \varnothing \\
\hline
\end{array}
$$

Test Plan Operations

Operations on test cases, test groups, and test plans are used to either define, update, delete or examine information. The need to create test information is clear. As the system evolves and undergoes changes existing tests may need to be modified or even deleted. To specify that an operation changes information the Δ notation is used :

$$
\begin{aligned}
\Delta \textit{TEST_CASE} &\ \widehat{=}\ [\ \textit{TEST_CASE},\ \textit{TEST_CASE}'\ ] \\
\Delta \textit{TEST_GROUP} &\ \widehat{=}\ [\ \textit{TEST_GROUP},\ \textit{TEST_GROUP}'\ ] \\
\Delta \textit{TEST_PLAN} &\ \widehat{=}\ [\ \textit{TEST_PLAN},\ \textit{TEST_PLAN}\ ]
\end{aligned}
$$

To specify that an operation does not change information the Ξ notation is used:

$$
\begin{aligned}
\Xi\ \textit{TEST_CASE} &\ \widehat{=}\ [\ \textit{TEST_CASE},\ \textit{TEST_CASE}'\ \mid \\
&\qquad \textit{TEST_CASE}' = \textit{TEST_CASE}] \\
\Xi\ \textit{TEST_GROUP} &\ \widehat{=}\ [\ \textit{TEST_GROUP},\ \textit{TEST_GROUP}'\ \mid \\
&\qquad \textit{TEST_GROUP}' = \textit{TEST_GROUP}\ ] \\
\Xi\ \textit{TEST_PLAN} &\ \widehat{=}\ [\ \textit{TEST_PLAN},\ \textit{TEST_PLAN}\ \mid \\
&\qquad \textit{TEST_PLAN}' = \textit{TEST_PLAN}\ ]
\end{aligned}
$$

Possible operations on test cases include:

- Create-tcase
- Change-tcase-label

- Change-tcase-item
- Set-tcase-input
- Set-tcase-result
- Set-tcase-procedure
- Get-tcase-label
- Get-tcase-item
- Get-tcase-input
- Get-tcase-result
- Get-tcase-procedure

Possible operations on test groups include:

- Create-tgroup
- Change-tgroup-label
- Set-tgroup-procedure
- Add-tgroup-test
- Delete-tgroup-test
- Get-tgroup-label
- Get-tgroup-procedure
- Get-tgroup-test
- List-tgroup-tests

Possible operations on a test plan include:

- Create-tplan
- Set-tplan-procedure
- Add-tplan-group
- Delete-tplan-group
- Get-tplan-group
- List-tplan-groups

Because the specification for many of these operations are similar only a few representative operations are defined below. Several operations on each object are given.

The operation *Create-tcase* is used to create a new test case. The operation only accepts a name for the test case and the item to be tested. Other information about the test case can be added using the set operations.

Create-tcase

$\Delta TEST_CASE$
$name? : LABEL$
$item? : TEST_ITEM$

$test' = name?$
$item' = item?$

$input' = \epsilon$
$result' = \epsilon$
$proc' = \epsilon$

The operation *Change-tcase-label* is used to change the label assigned to a test case.

Change-tcase-label

$\Delta TEST_CASE$
$new_label? : LABEL$

$test' = new_label?$

$item' = item$
$input' = input$
$result' = result$
$proc' = proc$

The operation *Set-tcase-input* specifies the input test data for the test case.

Set-tcase-input

$\Delta TEST_CASE$
$idata? : TEST_FORM$

$input' = idata?$

$test' = new_label?$
$item' = item$
$result' = result$
$proc' = proc$

The operation *Get-tcase-input* is used to retrieve the input test data for a test case.

Get-tcase-input

$\equiv$ *TEST_CASE*
idata! : *TEST_FORM*

---

idata! = *input*

The operation *Create-tgroup* creates a new test group. Only the name of the group is specified. The group test procedure and the individual test cases can be associated with the group later using the set operations.

Create-group

Δ *TEST_GROUP*
name? : *LABEL*

$name' = name?$

$proc' = \epsilon$
$tests' = \emptyset$

The operation *Add-tgroup-test* adds a test case to the group of test cases.

Add-tgroup-test

Δ *TEST_GROUP*
tc? : *TEST_CASE*

$tc? \notin tests$
$tests' = tests \cup \{tc?\}$

$name' = name$
$proc' = proc$

The operation *Delete-tgroup-test* removes a test case from the group of test cases.

Delete-tgroup-test

Δ *TEST_GROUP*
tc? : *TEST_CASE*

$tc? \in tests$
$tests' = tests - \{tc?\}$

$name' = name$
$proc' = proc$

The operation *Get-tgroup-test* retrieves a test case from the specified test case group.

Get-tgroup-test

$\equiv TEST_GROUP$
$tname? : LABEL$
$test! : TEST_CASE$

$\exists(l : LABEL,\ i : TEST_ITEM.\ in : TEST_FORM,\ out : TEST_FORM,$
$proc : PROC_NAME) \bullet (l, i, in, out, proc) \in tests \wedge \quad l = tname? \wedge$
$test! = (tname?, i, in, out, proc))$

The operation *List-tgroup-tests* returns a list of all the test case labels for a test group.

List-tgroup-tests

$\equiv TEST_GROUP$
$cases! : \mathbb{F}\ LABEL$

$cases! = \{ l : LABEL \mid \exists t : TEST_CASE \bullet t \in tests \wedge$
$Get\text{-}tcase\text{-}label(t) = l \}$

4.2.14 Estimates

Project planning requires the ability to accurately estimate the costs involved. A good understanding of what influences the cost is necessary to monitor, control, and in the future, reduce costs. These influences are often referred to as factors. Typical factors include system size, programming language(s), experience of the staff, and project methodology.

Estimation models make it possible to predict how much a project is likely to cost, the number of people needed, and the amount of effort. An estimation model has five major elements [Yu90].

Estimation Target : Identifies exactly what is to be estimated such as the project cost (measured in dollars), project effort (measured in staff-months), schedule or duration (measured in months).

Estimation Formulas : A set of algorithms (formulas) that produce the estimation target values by modeling the influence of various factors.

Estimation Process : The estimation process defines the standards and methods to assist with estimation. The process links estimation with other management activities, directs the collection of data used by the estimation model, and states when estimation formulas should be used.

Historical Project Database : A collection of information from completed projects with identified effort, staffing, and schedule results, as well as known environmental values.

Estimation Tools : As estimation model can be implemented in a tool that provides a user with a friendly way of accessing the historical database and generating requested estimates.

Before formalizing the concept of an estimation model, several sets must be defined. The set of estimation model targets is denoted by TARGET.

$$TARGET \mathrel{\widehat{=}} \{ \textbf{cost, effort, duration}, \ldots \}$$

The formulas used by an estimation model are equations. The set of all possible estimation model equations is denoted by EQUATION.

$$EQUATION \mathrel{\widehat{=}} \textit{The set of all possible estimation model equations}$$

The schema EST_MODEL denotes the information associated with an estimation model.

$$\begin{array}{|l}
\hline
\textit{EST_MODEL} \\
\textit{target} : TARGET \\
\textit{formulas} : \mathbb{F}\ EQUATION \\
\textit{process} : \mathbb{F}\{\ STANDARD_NAME \cup PROCEDURE_NAME\ \} \\
\textit{tools} : \mathbb{F}\ TOOL_NAME \\
\hline
\forall n : NAME \bullet n \in process \Rightarrow n \in \mathrm{dom}\, PROJECT\text{-}STANDARDS \vee \\
\qquad n \in \mathrm{dom}\, PROJECT\text{-}PROCEDURES \\
tools \subseteq PROJECT\text{-}TOOLS \\
\hline
\end{array}$$

The collection of estimation models used in a project is denoted by PROJECT-MODELS.

$$PROJECT\text{-}MODELS : \mathbb{F}\ EST_MODEL$$

Initially, there are no project models associated with a project. Though project planning is likely to quickly select one or more for use during the planning phase. The initial lack of project models is denoted by INIT-PROJECT-MODELS.

$$INIT\text{-}PROJECT\text{-}MODELS \mathrel{\widehat{=}} \varnothing$$

Note that a project database is not included in the estimation model information. There is only one historical project database and all estimation models use it. For now the nature of the historical project database will not be specified. The historical project database will simply be denoted by HIST_PROJ_DB.

One element of the historical project database is a collection of estimates. During the initial planning phase estimates for project cost, effort, and duration are made. As the project proceeds these estimates often need to be updated. Also estimates may be broken down by phase or even task. This estimation information can be formalized. The information for a single estimate is denoted by ESTIMATE.

$$\begin{array}{|l}
\hline
\textit{ESTIMATE} \\
\textit{type} : TARGET \\
\textit{item} : ITEM_NAME \\
\textit{date} : DATE \\
\textit{value} : \mathbb{Z} \\
\end{array}$$

The item element specifies what the estimate was for. For example, if the value of type is "effort" and item is "test system" then the estimate is for the amount of effort required to test the target system. The set of all possible item names is denoted by ITEM_NAME.

$$\begin{array}{|l} ITEM_NAME \mathrel{\widehat{=}} \textit{The set of all possible estimation item names} \\ \hline ITEM_NAME \subset NAME \end{array}$$

The collection of project estimates maintained in the historical project database is denoted by PROJECT-ESTIMATES.

$$PROJECT\text{-}ESTIMATES \mathrel{\widehat{=}} \mathbb{F}\ ESTIMATE$$

The initial contents of the historical project database depends on other projects that previously been recorded. At the beginning of a re-engineering project there are no estimates for the current project in the database base. This is denoted by INIT-PROJECT-ESTIMATES.

$$INIT\text{-}PROJECT\text{-}ESTIMATES = \varnothing$$

Estimate Operations

Necessary operations on estimation information include:

- Get-est-type
- Get-est-item
- Get-est-date
- Get-est-value

Necessary operations on project estimates include:

- Record-estimate
- Get-estimate
- Get-latest-estimate
- List-all-estimates
- List-latest-estimates

Necessary operations on estimation models include:

- Define-est-model
- Add-est-formula

- Update-est-formula
- Delete-est-formula
- Add-est-standard
- Delete-est-standard
- Add-est-procedure
- Delete-est-procedure
- Add-est-tool
- Delete-est-tool
- Get-est-target
- List-formulas
- List-est-standards
- List-est-procedures
- List-est-tools

Necessary operations on project estimation models include:

- Add-est-model
- Delete-est-model
- List-est-models

The specifications for these operations is not given.

4.2.15 Project Organizational Structure

Within a project there must be an organizational structure. At the top of the organizational structure is the project manager. Underneath this person are a number of teams each with an assigned role and responsibilities within the project. There may also be several people that serve as links between teams, or between teams and external groups such as the project client and users.

Once defined, the responsibility of the project organization is to work towards and achieve the successful planning, re-engineering, and accomplishment of the project goals. This is never an easy job and the job can be made harder or easier depending on the characteristics of the organizational structure and the key people within the project.

To formalize the concept of a project organizational structure, we begin by formalizing the concept of a role. Each project member fills one, possibly several, roles within a project. Also each team fills a role within a project. There are a number of different project roles including project

manager, team leader, coordinator, speciality staff, programmer, DBMS administrator, test group, support group, implementation group, etc. Each role has a title. Each role has an assigned group of responsibilities. Each role has a level of authority associated with it. Each role has a set of skills that a person or team fulfilling that role is expected to possess.

To formalize the concept of project roles several sets must be defined. The set of all possible skills is denoted by the set SKILL. The set of all possible authority levels is denoted by AUTHORITY. The set of all possible responsibilities is denoted by RESPONSIBILITY. The set of all possible titles is denoted by TITLE.

$SKILL \mathrel{\widehat{=}}$ *The set of all possible skills*
$AUTHORITY \mathrel{\widehat{=}}$ *The set of all possible authority levels*
$RESPONSIBILITY \mathrel{\widehat{=}}$ *The set of all possible responsibilities*
$TITLE \mathrel{\widehat{=}} \{$ **project-manager, team-leader, coordinator, programmer, DBMS-administrator, project-librarian, task-leader,** ... $\}$

A project role is denoted by the schema ROLE.

```
┌─ ROLE ──────────────────────────────────────────
│ title : TITLE
│ duties : F RESPONSIBILITY
│ auth : F AUTHORITY
│ skills : F SKILLS
└─────────────────────────────────────────────────
```

At the lowest level of detail within the organizational structure information about project members is given. Each member has a name. Each member has a date that they are expected to start work on the project and a date after which they are expected to be re-assigned off the project. Such information is useful for scheduling multiple projects within an organization. Each member has one or more tasks to which they are assigned. Each member is assigned to a team. One issue is whether a person can serve on more than one team. Each person has an assigned role within a team. Thus a person assigned to multiple teams may have multiple roles to perform. Information about a project member is formalized by the schema MEMBER.

```
┌─ MEMBER ────────────────────────────────────────
│ start_date : DATE
│ stop_date : DATE
│ assigned_to : TEAM_NAME ⇸ ROLE
│ assignments : F TASK_NAME
└─────────────────────────────────────────────────
```

When a new member is initially assigned to a project there may be no information about that member yet. Therefore, a constant schema NEW-MEMBER is defined that represents initial member information.

$$
\begin{array}{|l}
\hline
\textit{NEW-MEMBER} \\
\textit{MEMBER} \\
\hline
\textit{start_date} = \epsilon \\
\textit{stop_date} = \epsilon \\
\textit{assign_to} = \varnothing \\
\textit{assignments} = \varnothing \\
\hline
\end{array}
$$

The list of all personnel assigned to a project is denoted by PROJECT-PERSONNEL.

$$\textit{PROJECT-PERSONNEL} : \textit{PERSON_NAME} \nrightarrow \textit{MEMBER}$$

Project members are assigned to teams. A team is the focal point for achieving task goals. Each team is assigned one or more tasks and is responsible for accomplishing those tasks efficiently, on-time, and according to the task procedure (task plan). Teams map closely to tasks. Each team has a team leader who is also the leader of at least one task. Workers assigned to tasks lead by this person are drawn from this person's team. Note it is possible to assign members of different teams to a task. For example, a review task often requires people from different teams. The concept of a team is denoted by TEAM.

$$
\begin{array}{|l}
\hline
\textit{TEAM} \\
\textit{leader} : \textit{PERSON_NAME} \\
\textit{members} : \mathbb{F}\ \textit{PERSON_NAME} \\
\textit{tasks} : \mathbb{F}\ \textit{TASK_NAME} \\
\textit{role} : \textit{ROLE} \\
\hline
\end{array}
$$

Information about project teams is denoted by the function PROJECT-TEAMS.

$$\textit{PROJECT-TEAMS} : \textit{TEAM_NAME} \rightarrow \textit{TEAM}$$

Where the set TEAM_NAME is defined as:

$$
\begin{array}{|l}
\textit{TEAM_NAME} \mathrel{\widehat{=}} \textit{The set of all possible team names} \\
\hline
\textit{TEAM_NAME} \subset \textit{NAME}
\end{array}
$$

Organizational Structure Operations

There are several operations that apply to role information. These operations include:

- Create-role
- Add-role-duty
- Add-role-authority
- Add-role-skill

- Delete-role-duty
- Delete-role-authority
- Delete-role-skill
- List-role-duty
- List-role-authority
- List-role-skill

There are several operations that can be defined on member information. These include:

- Set-memb-start
- Set-memb-stop
- Assign-memb-team
- Assign-memb-task
- Remove-memb-team
- Remove-memb-task
- List-memb-teams
- List-memb-roles
- List-memb-team-role
- List-memb-tasks

There are several operations that can be defined on project personnel information. These include:

- Add-proj-member
- Delete-proj-member
- Get-proj-member
- Update-proj-member
- List-personnel

There are several operations that can be defined on team information. These include:

- Set-team-leader
- Add-team-member

- Add-team-task
- Delete-team-member
- Delete-team-task
- Get-team-leader
- List-team-members
- List-team-tasks

There are several operations that can be defined on project team information. These include:

- Add-proj-team
- Delete-proj-team
- Get-proj-team
- Update-proj-team
- List-proj-teams

To denote that an operation changes information the Δ notation is used.

$$\Delta ROLE \mathrel{\widehat{=}} [\, ROLE, ROLE' \,]$$
$$\Delta MEMBER \mathrel{\widehat{=}} [\, MEMBER, MEMBER' \,]$$
$$\Delta PROJECT\text{-}PERSONNEL \mathrel{\widehat{=}} [\, PROJECT\text{-}PERSONNEL, PROJECT\text{-}PERSONNEL' \,]$$
$$\Delta PROJECT\text{-}TEAMS \mathrel{\widehat{=}} [\, PROJECT\text{-}TEAMS, PROJECT\text{-}TEAMS' \,]$$

To denote that an operation does not change information the Ξ notation is used.

$$\Xi ROLE \mathrel{\widehat{=}} [\, ROLE, ROLE' \mid ROLE' = ROLE \,]$$
$$\Xi MEMBER \mathrel{\widehat{=}} [\, MEMBER, MEMBER' \mid MEMBER' = MEMBER \,]$$
$$\Xi PROJECT\text{-}PERSONNEL \mathrel{\widehat{=}} [\, PROJECT\text{-}PERSONNEL, PROJECT\text{-}PERSONNEL' \mid PROJECT\text{-}PERSONNEL' = PROJECT\text{-}PERSONNEL \,]$$
$$\Xi PROJECT\text{-}TEAMS \mathrel{\widehat{=}} [\, PROJECT\text{-}TEAMS, PROJECT\text{-}TEAMS' \mid PROJECT\text{-}TEAMS' = PROJECT\text{-}TEAMS \,]$$

The operation *Create-role* is used to create a new role object to hold information about somebody or some team. At the time of creation only the role title is set. The remainder of the role information must be filled in using the add-role operations.

Create-role

$$\Delta ROLE$$
$$title? : TITLE$$

$$title' = title?$$
$$duties' = \varnothing$$
$$auth' = \varnothing$$
$$skills' = \varnothing$$

The operation *Add-role-duty* is used to add a new responsibility to an existing role.

Add-role-duty

$\Delta ROLE$
$new_duty? : RESPONSIBILITY$

$new_duty? \notin duties$
$duties' = duties \cup \{ new_duty? \}$

$title' = title$
$auth' = auth$
$skills' = skills$

The operation *Delete-role-duty* is used to remove an assigned responsibility from a role.

Delete-role-duty

$\Delta ROLE$
$ex_duty? : RESPONSIBILITY$

$ex_duty? \in duties$
$duties' = duties - \{ ex_duty? \}$

$title' = title$
$auth' = auth$
$skills' = skills$

The operation *List-role-duty* is used to list all the responsibilities assigned to the specified role.

List-role-duty

$\Xi ROLE$
$list! : \mathbb{F}\, RESPONSIBILITY$

$list! = duties$

The other operations on role information that manipulate the role's authorities and skills are similar to the duty operations. These other operations are not given here.

The operation *Set-memb-start* is used to set the starting date for a person assigned to a project. This is the date on which the person is to start work on the project.

Set-memb-start

$\Delta MEMBER$
$start? : DATE$

$start_date' = start?$

$stop_date' = stop_date$
$assigned_to' = assigned_to$
$assignments' = assignments$

The operation *Set-memb-stop* specifies the date on which the person is to stop working on the project.

Set-memb-stop

$\Delta MEMBER$
$stop? : DATE$

$stop_date' = stop?$

$start_date' = start_date$
$assigned_to' = assigned_to$
$assignments' = assignments$

The operation *Assign-memb-team* is used to assign a person to a project team. Note that it is possible for one person to be assigned to more than one team. Whether this is a good idea or not is up to management. When assigning a person to a team that person must also be given a role in the team.

Assign-memb-team

$\Delta MEMBER$
$team? : TEAM_NAME$
$role? : ROLE$
$\Xi PROJECT\text{-}TEAMS$

$team? \in \mathrm{dom}\, PROJECT\text{-}TEAMS$
$team? \notin \mathrm{dom}\, assigned_to$
$assigned_to' = assigned_to \cup \{ team? \mapsto role? \}$

$start_date' = start_date$
$stop_date' = stop_date$
$assignments' = assignments$

The operation *Assign-memb-task* is used to assign a person to a task.

Assign-memb-task

$\Delta MEMBER$
$task? : TASK_NAME$
$\Xi PROJECT\text{-}TASK$

$task? \in \mathrm{dom}\, PROJECT\text{-}TASK$
$task? \notin assignments$
$assignments' = assignments \cup \{ task? \}$

$start_date' = start_date$
$stop_date' = stop_date$
$assigned_to' = assigned_to$

The operation *Remove-memb-team* is used to take a person off a team to which the person is assigned.

Remove-memb-team

$\Delta MEMBER$
$team? : TEAM_NAME$
$\Xi PROJECT\text{-}TEAMS$

$team? \in \mathrm{dom}\, PROJECT\text{-}TEAMS$
$team? \in \mathrm{dom}\, assigned_to$
$assigned_to' = \{ team? \} \lhd assigned_to$

$start_date' = start_date$
$stop_date = stop_date$
$assignments' = assignments$

Other operations on the organizational structure information are not given at this time.

4.2.16 Scheduling

Every project has a schedule that specifies when a task will begin and when it is expected to end. At the beginning of a project an initial schedule is developed and as the project progresses the schedule must often be changed. The schedule is closely watched by project management and a great deal of importance is frequently given to scheduling tasks such that the promised due date can be met.

There is no separate schedule object within this specification. Information used by a schedule are task names and planned start and stop dates. The start and stop information for each task is held within the task information as formalized by TASK. By defining a schedule the planning information within each task is simply updated.

Schedule Operations

While there is no schedule object there are several operations on the schedule information that are necessary. These operations include:

- Set-sched-dates
- Change-sched-start
- Change-sched-stop
- List-schedule

In the specification of these operations that follow, one disadvantage that occurs when a new date is specified, is that the currently specified date is lost. During project tracking the slippage of scheduled dates should be noted and recorded. This shortcoming can be corrected by modifying the definition of the TASK object and these operations. However, for now the shortcoming is left in.

The operation *Set-sched-dates* is used to initially set the start and stop dates for a single specified task.

Set-sched-dates

$\Delta PROJECT\text{-}TASK$
$task? : TASK_NAME$
$start? : DATE$
$stop? : DATE$

$PROJECT\text{-}TASK' = PROJECT\text{-}TASK \oplus \{ task? \mapsto$
$\quad Assign\text{-}task\text{-}dates(PROJECT\text{-}TASK(task?), start?, stop?) \}$

The operation *Change-sched-start* is used to update just the planned starting date assigned for a task.

Change-sched-start

$\Delta PROJECT\text{-}TASK$
$task? : TASK_NAME$
$start? : DATE$

$PROJECT\text{-}TASK' = PROJECT\text{-}TASK \oplus \{ task? \mapsto$
$\quad Replace\text{-}task\text{-}pstart(PROJECT\text{-}TASK(task?), start?) \}$

The operation *Change-sched-stop* is used to update just the planned stopping date assigned to a task.

Change-sched-start

$\Delta PROJECT\text{-}TASK$
$task? : TASK_NAME$
$stop? : DATE$

$PROJECT\text{-}TASK' = PROJECT\text{-}TASK \oplus \{ task? \mapsto$
$\quad Replace\text{-}task\text{-}pstop(PROJECT\text{-}TASK(task?), stop?) \}$

The operation *List-schedule* is used to generate a report on the project schedule. Here a relation is created that specifies an ordering among project tasks.

List-schedule

$\Xi PROJECT\text{-}TASK$
$schedule! : TASK_NAME \leftrightarrow TASK_NAME$

$schedule! = \{ n_1, n_2 : TASK_NAME \mid n_1 \in \mathrm{dom}\, PROJECT\text{-}TASK \wedge$
$\quad n_2 \in \mathrm{dom}\, PROJECT\text{-}TASK \wedge$
$\quad (Get\text{-}task\text{-}pstart(n_1) = Get\text{-}task\text{-}pstart(n_2) \vee$
$\quad\quad Get\text{-}task\text{-}pstart(n_1)\, AFTER\, Get\text{-}task\text{-}pstart(n_2)) \}$

4.3 Tasks

There are many tasks that must be done to prepare a re-engineering project. These tasks require input from the outside world. These tasks in turn create new information and knowledge that is used during a project. The information objects and the operations defined in the previous section are used by these tasks.

Much of the input to the Analysis and Planning phase is human knowledge and information. However, there are several other inputs as well. These other inputs include:

Input

- Existing System Source Code
- Existing System Documentation
- Existing System Test Plans
- PROJECT-MODELS
- HIST_PROJ_DB

There are many outputs from this task as well. Each of the outputs listed below is created by a sub-task of the Analysis and Planning phase.

Output

- ACCEPT_CRITERIA
- DEFINITION
- ENVIRONMENTS
- INVENTORY
- RESOURCES
- PROJECT-ESTIMATES
- PROJECT-PERSONNEL
- PROJECT-PLAN
- PROJECT-PROCEDURES
- PROJECT-STANDARDS
- PROJECT-TASKS
- PROJECT-TEAMS
- PROJECT-TOOLS
- TEST-PLAN

The sub-tasks of the Analysis and Planning phase are discussed in the following subsections. Remember the Analysis and Planning phase is a task itself. For each subtask, a description of the task, a list of inputs and outputs is given, followed by a discussion of concerns or processing within the task. These are not complete, in-depth descriptions of these subtasks. The complete details of each task is too voluminous to be given here.

4.3.1 Determine Motivations and Objectives

This is the first task of the Analysis and Planning phase. The purpose of this task is to identify the motivations for re-engineering the existing system and the general project objectives. This task focuses on two questions. *Why does the system need to be re-engineered? What is this project supposed to achieve?* The answers to these questions can be used to determine the scope of the project. This allows the project planners to state what can and can not be done during the project.

Input

- Application Domain Knowledge
- Environment Knowledge
- Existing System Knowledge
- Programming Language Knowledge
- Software Re-engineering Knowledge

Output

- DEFINITION

DISCUSSION :

The first step in this task is to identify the motivations for the re-engineering project. Either there are problems with the system that must be addressed or there are changes in the system's environment that affect the system. Perhaps there are new requirements that need to be addressed. Future plans for the system should also be considered and preparations incorporated into the system.

The second step is to identify the project objectives. The objectives describe what the re-engineering project is to achieve, and in general terms, describe the desired characteristics of the target system. For each project motivation there is an objective to satisfy that motivation. In addition, by deciding to embark on a re-engineering project other objectives that are not related to the original motivations are often included.

4.3.2 Analyze Environments

This task is only necessary if a motivation for the re-engineering project is to convert the system to operate in a new environment. The new environment is called the target environment. In the lesser case the target environment results from an upgrade or minor change in the existing environment, in the more extreme case the existing system is to be ported to a new operating system or computer. Here, software conversion [OSD/FSMSC89] becomes part of the re-engineering project.

The goal of this task is to identify differences in the existing and target environments that can affect the system. For example, if the system reads binary encoded files then word size, byte order, and data representations must be examined. There may be differences in operating system utilities that the system uses. Any difference in the environments that affects the system will have to be handled during the re-engineering project. Better to know about the differences in advance than to stumble across them during re-implementation or testing.

Input

- DEFINITION
- Environment Knowledge
- Programming Language Knowledge
- Re-engineering Knowledge
- System Knowledge

Output

- ENVIRONMENTS

The first step in this task is to determine if the project is motivated by a change in the existing environment. If the existing and target environments are the same then this task is not necessary. If there have been (or will be) changes in the existing environment or the system is being ported to a new environment then this step is necessary.

If there is a change in the environment the second step is to analyze the existing environment and list its properties and characteristics. In practice, the system will be examined and possible environment properties that may affect the system operation identified. This list is used to guide the environment analysis. Though it is likely to not be complete so other properties of the environment should still be determined.

The third step is to analyze the target system. Again in practice, the list of environment properties that could affect the system will be used to guide the analysis. Experience in software conversions will also help guide the analysis.

The fourth step is to compare the results of the two analyses and identify important differences. The comparison looks for several types of differences. If the system uses a feature of the existing environment and that feature is different in the target environment that the difference must be noted. If the system uses a feature in of the existing environment and that feature is not present in the target environment that difference must be noted. This is actually, the most difficult case, since the missing feature must be developed on the target environment. Finally, any features in the target environment that maybe useful and that are not found in the existing environment should be noted. This is particularly true when the target environment is an upgrade of the existing environment.

4.3.3 Inventory Task

The purpose of the inventory task is to identify all objects that make up the system and all the supporting objects that are needed. These are objects that will either be re-engineered or used during the project. A system is composed of programs, JCL commands, data files, and databases. User created or vendor supplied libraries may also be used in system programs. In addition, there may be system documentation, test plans, test cases, test data files, etc, that must be identified.

Input

- DEFINITION
- Cost Estimation Knowledge
- Environment Knowledge
- Existing Documents
- Existing System Knowledge
- Programming Language Knowledge
- Re-engineering Knowledge

Output

- INVENTORY

DISCUSSION :

The inventory task could follow this sequence of steps.

1. Use the project objectives to identify the system being re-engineered.
2. Use the system knowledge to identify all objects in the system. This includes supporting objects.
3. Use environment knowledge and system knowledge to locate objects.
4. Use estimation knowledge, system knowledge, environment knowledge, programming language knowledge, project objectives, and re-engineering knowledge to determine the information to be collected for each object.
5. Use system knowledge, environment knowledge, and programming language knowledge, estimation process knowledge, file names, file names, and existing documentation to collect the required object information.

The first problem is to determine what is needed. If the entire system is to be re-engineered then information about the entire system must be collected. If only parts of the system are to be re-engineered then these parts must be identified and information about them collected. The project objectives will help clarify what is to be affected. If the entire system is not to be re-engineered then parts of the system that will not be changed but which are coupled with the parts to be re-engineered must be identified also.

Another problem is if multiple versions of the system exist. A re-engineering goal may be to combine the different versions. Here, it will be necessary to inventory the separate versions of the system.

Documents for the system must be collected. Some documents will only exist as hardcopies, others will have parts or the entire text stored in computer files. A possible project goal might be to update or replace documents. The quality of the existing documents must be determined.

Besides identifying the system objects it is necessary to collect information about each object. The necessary information will differ from project to project and from organization to organization. At the least, information required to make estimates of project cost and effort must be

collected. Knowledge about the estimation models and their required inputs will help determine the information to be collected.

Techniques are also required to collect information. The estimation process may specify techniques and metrics to use to collect information.

The inventory task uses a wide variety of inputs. Knowledge about the system is necessary to identify its components. Operations for manipulating the inventory information have been specified. Operations and procedures for collecting information, and for identifying objects are also needed by this task.

4.3.4 Define Approach

This task is an important and time-consuming task within the Analysis and Planning phase. This task focuses on one question: Given the current software system and the properties of the target system how can the target system be created? The answer to this question is the project plan.

Input

- ENVIRONMENTS
- DEFINITION
- PROJECT-PROCEDURES
- Application domain knowledge
- Existing system knowledge
- Management knowledge
- Re-engineering knowledge
- Software engineering knowledge

Output

- PROJECT-PLAN
- PROJECT-TASK

DISCUSSION:

The major phases of a re-engineering project are given by re-engineering knowledge. However, this does not simplify the job of organizing and planning for each phase. One difficult question that must be addressed is whether the system is to be re-engineered in one "lump sum," in fragments, or incrementally.

Sometimes a system is just too large to completely re-engineer at one time. Doing so many result in a long duration project. Sometimes it is better to divide the system into fragments where each fragment can be re-engineered separately. The problem with this approach is identifying fragments that can be isolated from the rest of the system. An alternative is to re-engineer the system incrementally. This involves setting a series of minor goals and re-engineering the system

to met these goals one increment at a time. The result of each increment is an improved system that can be used while the next increment is under work.

A combination of re-engineering knowledge, project objectives, and existing system knowledge must be used to identify how the project will be conducted. Besides establishing the grand or macro plan that identifies major phases and tasks, it is necessary to establish micro plans, specifying the details of each major task. Re-engineering knowledge will identify useful tasks, project objectives will enable the selection of necessary tasks, and existing system knowledge will determine the details of how the system can be re-engineered.

4.3.5 Define Acceptance Criteria

The acceptance criteria are used during acceptance testing to verify that the target system is acceptable to the client. The acceptance criteria must be agreed on early in the project, typically before the technical work begins.

Input

- DEFINITION
- Environment knowledge
- Existing system knowledge
- Application domain knowledge

Output

- ACCEPT_CRITERIA

DISCUSSION :

The criteria specified must be specific. Each criterion must be verifiable. A criterion such as, "The system will operate correctly on valid inputs." is not a good criterion since this typically can not be proven for a large system.

Acceptance criteria should complement, not duplicate the internal project quality assurance procedures. Of course, one criterion can be that the system successfully pass all unit and system level tests.

4.3.6 Define Project Procedures

The purpose of the project procedures task is to assemble the set of procedures to be used during a re-engineering project. The organization may have a collection of standard procedures that must be followed. Some of these procedures may need to be adjusted to fit specific characteristics of the current project and its environment. Each project is different, so a certain amount of adjustment of procedures is to be expected. New procedures may also need to be developed to handle tasks unique to re-engineering or the current project.

Input

- See Below

Output

- PROJECT-PROCEDURES

DISCUSSION : The input for this task was left unspecified. The reason for this is that to collect and tailor existing procedures knowledge of each procedure and its domain is required. Also knowledge of the specific project details. To develop new procedures knowledge of the procedure domains is again needed. To collect, tailor, and develop a complete and useful set of procedures for a project, complete knowledge of every aspect of the project and work required is necessary.

4.3.7 Define Project Standards

The purpose of the standards task is to assemble the set of standards that will be enforced during the project.

Input

- See Below.

Output

- PROJECT-STANDARDS

DISCUSSION :

The input for this task was left unspecified. The reason for this is the broad range of knowledge sources that must be used to examine, select, define, and change the standards.

This task will typically begin by collecting existing standards. The initial project objectives will help identify useful and/or necessary standards. Some of these may need to be modified to conform with project specific requirements. An examination of the proposed project plan will also be helpful in identifying standards that are needed may not exist within the organization. For such standards it will be necessary to define new standards.

4.3.8 Identify Resources

Part of the task of planning a project is identifying any special resources that a project will require. The tasks that require these resources should be noted, since the availability of a resource can affect the project schedule. The need for a resource may be typical of the system or may become apparent as particular tasks are planned.

Input

- Existing System Knowledge

- Environment Knowledge
- Organization Knowledge
- Re-engineering Knowledge
- Tool Knowledge

Output

- RESOURCES

DISCUSSION :

While defining the re-engineering environment special equipment needs may be identified. This can range from computers to be used to re-engineering the code, such a workstations, to special purpose hardware used by the system. The need for special purpose hardware can affect the system tests and system transition plans.

Resources can be identify by considering several factors:

- Does the system required special purpose hardware to operate? For example, the software that drives a HUD display will need a HUD hardware device. Another example, is the need for a color graphics system to use a color display device.
- Does the system require special purpose hardware for testing?
- Does equipment need to be allocated to the project such as computers, terminals, or printers?
- Does the system require special capabilities such as access to a network?
- Are there key people associated with the system or the environment who will not be assigned to this project, but whose knowledge and expertise is important?

The answers to these questions and others will identify the resources required by a project. Identifying needed resources early in the project simplifies the task of obtaining the resources before they are needed.

4.3.9 Identify Tools

The task of identifying tools to be used during a project is likely to span almost the entire planning stage. As work proceeds in other planning tasks, tools will be identified and added to the list of project tools.

Input

- PROJECT-PROCEDURES
- PROJECT-TASKS
- Environment knowledge

- Management knowledge
- Re-engineering knowledge
- Testing knowledge
- Tool knowledge

Output

- PROJECT-TOOLS

DISCUSSION :

The selection of the appropriate tools can have a great effect on the cost and duration of a project. Some tools are taken for granted such as text editors and language compilers. Other tools such as configuration managers, document management systems, software engineering environments, and test case generators are not used in all organizations. The value of some tools may not be appreciated by those who have not used them before or by those who have not confronted the problems such tools address.

One problem that a project may face is that a desired tool does not exist in-house. The alternatives are to purchase the tool, develop it in-house, or do without. The pro's and con's of each of these alternatives must be judged carefully as the decision can have an effect on the project schedule.

4.3.10 Test Planning

This task focuses on preparing the test plan for the target system. It may not be possible to specify the details of each individual test during the planning stage, but it is possible to plan the testing strategy. Test planning for a re-engineering project differs from test planning for software development since there may already be an existing test plan for the existing system. Work on the target system test plan can begin by examining the existing system test plan.

Input

- INVENTORY
- DEFINITION
- ENVIRONMENTS
- Application domain knowledge
- Existing system knowledge
- Existing test plan
- Programming language knowledge
- Tool knowledge

Output

- TEST_PLAN

DISCUSSION :

The first step is to identify a testing strategy for the target system. Will white-box or black-box or both strategies be used. What percentage of coverage will be required of white-box tests? Will automatic test case generators be used? How will the system be tested? The answers to these questions will be used to guide the development of a test plan for the target system.

The second step is to analyze the existing system test plan and determine its suitability for testing the target system. The focus of this step is to determine which parts of the existing test plan can be re-used without changes, re-used with changes, and which parts must be deleted. The existing test plan is compared against the planned changes to the target system, this can be done using the project DEFINITION information. For example, if the existing test cases are based on white-box testing and a project goal is to re-design and re-implement a portion of the system, then the test cases that exercise that portion of the system can not be re-used. Existing test cases must be modified or changed if they exercise features of the existing system (black-box tests) that will be changed during the project.

The third step is to identify portions of the target system for which no existing tests will exercise. For example, new features to be added to the target system, or code that will be replaced and is currently tested by white-box based tests. Plans for testing these portions of the target system must be made.

Notice that tests that are designed to cover target system features can be generated during the planning stage once the requirements for the target system are complete. White-box test cases that are designed to cover statements or branches in the target system source code can not be generated until later in the project when the target source code exists. However, plans for testing such sections of the system can be made during the planning stage.

The product of this task is a test plan for the target system. The test plan may contain test groups for unit and system tests. To create the test plan an understanding of system testing, the existing system, the target system, and changes between the two is necessary. To specify details of the test cases knowledge about how the tests will be applied is also required.

4.3.11 Estimation

The key question behind a re-engineering estimation model is: Given an existing software system how much effort will it take to generate the desired target system? Thus, estimation models for software re-engineering are characterized by the concept of change. A fundamental assumption underlying software engineering is that an existing software system with undesirable properties is altered to create a new system that is somehow better. To estimate the effort required to re-engineer an existing software system it will be necessary to estimate the degree of change the existing system must undergo. The greater the expected amount of change the greater the effort required to produce the change.

Input

- DEFINITION
- ENVIRONMENTS

- PROJECT-TOOL
- PROJECT-MODELS
- HIST_PROJ_DB
- Re-engineering knowledge
- Management knowledge

Output

- PROJECT-ESTIMATES

DISCUSSION :

The concept of change can be embodied in a top-down estimation model for software re-engineering. A top-down model focuses on the re-engineering product, which is typically source code (it can also be documentation or test plans). A top-down model must include at least three components. The first component is some measure of the properties of the existing system. The second component is some measure of the properties of the target system. These two measures can be used to estimate the expected degree of change. The third component is some measure of the influence exerted by various factors.

The concept of change can also be embodied in a bottom-up estimation model for software re-engineering. A bottom-up model focuses on the re-engineering process. Software re-engineering is based on three fundamental principles: abstraction, alteration, and refinement. These principles are implemented in the re-engineering process by the technical phases: reverse engineering, re-design, and re-implementation. The expected degree of change affects each of these major phases. The portions of the system affected by the proposed changes must be reverse engineered to a suitable level of abstraction. The expected degree of change will affect the amount of work done in the re-design phase. Finally, all portions of the system that undergo change must be re-implemented.

4.3.12 Define Organizational Structure

The organizational structure of a project must be assembled carefully. It is more than just assigning the right people to the right jobs. The structure will define the flow of work between teams. There are several possible structures:

Assembly line Divides the project into functional teams. There is a management team, a reverse engineering team, a quality assurance team, an implementation team, etc. Each team works only in its functional area.

Parallel teams Each team performs similar tasks, but work on different portions of the system. Thus one team will handle tasks from many functional areas. This approach often causes better project morale since the work assignments vary for a person, but this approach requires a longer learning curve.

Input

- PROJECT-TASK
- Management knowledge
- Re-engineering knowledge
- Organization knowledge

Output

- PROJECT-TEAMS
- PROJECT-PERSONNEL

DISCUSSION :

The organizational structure and the size of the teams may vary from phase to phase. Factors that influence the structure include the staffing and resource requirements of the tasks, team functions, number and type of tasks, and the skill level of the personnel.

Initially the number of people assigned to a project will be small. The number will increase as the planning phase finishs and the other phases begin. It is important that the key project people be assigned to the project at the beginning or as early as possible. These people must have the best understanding of the project and what it involves. This knowledge can best be gained by participating in the project planning. Another important point is that project personnel must be assigned to the project full time. Re-engineering is not a part-time job.

4.3.13 Schedule Tasks

Developing a good project schedule means solving a difficult puzzle. Estimating how long a task will take to complete is difficult. Typically, a project schedule will be revised several times during a project to reflect slippage in the schedule.

Input

- PROJECT-TASK
- Management knowledge
- Re-engineering knowledge
- Tool knowledge
- Organization knowledge

Output

- PROJECT-TASK'

DISCUSSION :

To adequately schedule the project tasks an understanding of what each task involves is required. The tools available for a task and the people assigned to a task are also factors that must be

considered. Personnel may have learning curves to climb either in doing re-engineering work or using new tools, or a new environment. This lengthens the task time.

Care must be given to the ordering of tasks. Bottlenecks must be detected and if possible removed so that portions of a project don't come to a halt while waiting for other tasks to finish. There are a number of commercial scheduling tools available. While these tools have been developed to schedule software development projects, there is no real difference between the scheduling needs of software development and software re-engineering, though the technical work is clearly different.

4.3.14 Review Project Plans

The final task of the Analysis and Planning phase is to review the information collected during this phase. The project approach, procedures, objectives, acceptance criteria, etc, must be reviewed and accepted before the next phase of the project can begin.

Input

- ACCEPT_CRITERIA
- DEFINITION
- ENVIRONMENTS
- INVENTORY
- RESOURCES
- PROJECT-ESTIMATES
- PROJECT-PERSONNEL
- PROJECT-PLAN
- PROJECT-PROCEDURES
- PROJECT-STANDARDS
- PROJECT-TASKS
- PROJECT-TEAMS
- PROJECT-TOOLS
- TEST-PLAN

Output

- PROJECT-TASKS

DISCUSSION :

Before the software re-enginering project can continue the project plan must be reviewed and accepted not only by project management, but by upper management as well. The support of upper management is important to a re-engineering project and upper management must be aware of what such a project involves. Try to avoid surprises down the road.

If errors are found in the project information, or potential problems are spotted, it may be necessary to go back and repeat one or more analysis and planning tasks to correct the problems. Potential problems should not be passed on with the attitude that they will be dealt with when they arise. It is better to resolve problems before they become problems!

When the project plan is accepted, authority to begin the next phase of the project is given. This causes an update in the status of the project which is why PROJECT-TASKS is listed as an output of this task.

Chapter 5

Conclusions

An issue that arises while modeling and describing the software re-engineering process is whether any such model is valid. To model the process is to write a description of how software re-engineering is done. The sources of information used to create a process model must be evaluated. One source of information is experience taken from successful and unsuccessful re-engineering projects. This results in a description of how the process *can* be done. Another source is conjecture based on the modeler's understanding of re-engineering. This results is a description of how the process ***could*** be done.

A process modeler faced with a poorly understood portion of a process may create a model that seems reasonable but is not based on re-engineering experience. Here, the modeler creates a possible model of the process. This is a plausible process model, one that is not based on proven experience. A plausible process model describes an untried process. A plausible model may not be suitable for guiding a real project, i.e. it may not allow valid actions within the process or may omit necessary actions.

A process model based on re-engineering experience has more validatity, but only so far as it captures a true understanding of the dynamics of the re-engineering process. Experience must be accumulated from many projects. Each project is different. It is necessary to abstract common process elements from each project. When reviewing multiple projects it is easy to focus on some aspects of the process and ignore others. This leads to an incomplete process model.

A high-level process model is emerging from reported software re-engineering experience and research. Unfortunately, reports on re-engineering projects tend to focus on the technical aspects of the projects. This has resulted in a lack of reported information about other aspects of re-engineering projects such as management and support concerns.

One of the contributions of this work is a more complete model of the software re-engineering process. Chapter 3 presented a high-level process model for software re-engineering that extends the traditional re-engineering process model by including other phases that should occur within software re-engineering projects. The traditional model is derived from reported experience with software re-engineering. The extented process model presented in Chapter 3 is a plausible process model. The basis for the extensions derives from experience with software development and software conversion. There is more to software re-engineering that just technical work and the extented process model

addresses these previously ignored phases.

The extensions to the traditional process model create a more complete model of the software re-engineering process. It is apparent from studying the model that there is a broad range of issues and concerns that must be addressed to adequately capture and model the re-engineering process. There are many details to be addressed within re-engineering, far more than just the technical concerns.

The specification of the Analysis and Planning phase given in Chapter 4 demonstrates the range of concerns and volume of information that must be modeled. This specification focused on only one phase, and within that only focused on information and operations on this information. There are numerous details about this phase that are not covered in the specification. The Analysis and Planning phase has many tasks that must be specified in greater detail. The specification of this phase is a good start, but there is much work left to be done. The goal is to describe the process in enough detail that the model can be used to successfully guide a re-enginering project.

There are several difficulties that prevent the quick specification of software re-engineering process details. First, software re-engineering has not received much attention until lately. Our understanding of the problem and possible solutions is growing. Software development has been addressed for over 30 years and it is still not possible to complete specify the development process. A detailed model of the software re-engineering process will take time. First, more experience with re-engineering must be gained.

Software re-engineering project details must be reported. Re-engineering projects must be tracked and detailed information about the processes used within a project must be recorded. This information must be shared so that common process elements and approachs can be identified. A good high-level model of the re-engineering process is emerging. If the model is to be filled out to lower levels of detail then experience must be collected and shared.

The software re-engineering process is complex. Given the range of concerns and issues within re-engineering it may be better to focus on individual tasks within the process and specify these. This means solving the problem a piece at a time. Eventually, the pieces can be assembled to create a complete process model.

The high-level process model is needed to provide a framework for understanding and describing the process. A high-level process, such as the one presented in Chapter 3, identifies the pieces that must be investigated. Even the specification in Chapter 4, which fills in many details about one phase of the process, reveals tasks that require more indepth research and modeling.

Some software re-engineering tasks are similar to tasks in other domains such as software development and software conversion. There is a wealth of detailed experience in software development. By identifying similarities between software development tasks and software re-engineering tasks additional process details can be identified. For example, task scheduling, the schedule for a re-engineering project is just as important as the schedule for a software development project. There is a great deal of research results, experience, and tools available that address scheduling software development tasks, much of this is applicable to software re-engineering as well.

The current approach to modeling the software re-engineering process is to continue to evolve the high-level model and learn where more process details are necessary. By studying re-engineering projects additional process details will be detected. By comparing software re-engineering and soft-

ware development, experience from development can be reused to further describe the re-engineering process. Finally, future research can investigate better means of re-engineering software, leading to more useful models of the re-engineering process.

It must be realized that there is no one software re-engineering process model. Every project is different, just as different approachs are used to develop software for business, scientific, and military applications, so are different re-engineering approachs needed also. Re-engineering process models must be able to address different re-engineering approachs.

One re-engineering approach is **lump sum** re-engineering. Here, an entire software system is re-engineered at the same time, starting with the existing system and ending with the desired target system. However, this approach is not always suitable. One reason is the drain on resources to re-engineer a large system.

Another re-engineering approach is **incremental re-engineering**. Here, an existing system is re-engineered in increments. The entire system is re-engineered, but instead of producing the desired target system, one or more intermediate systems are produced. Each intermediate svstem has better properties than the preceding system version and is closer to the desired target system. This approach has the advantage that improved versions of the system are produced more quickly, but the length of time required to produce the desired target system may be longer than with lump sum re-engineering.

A third re-engineering approach is **partial re-engineering**. Here, only a portion of a system is re-engineered instead of the entire system. This places constraints on the resulting target system, since the re-engineered code must integrate with the existing non re-engineered system code. The range of flexibility to redesign the system is reduced with this approach.

Another issue in modeling the re-engineering process is how to incorporate requirement changes into a software system. Software re-engineering is a change in the *form* of a software system, not a change in *functionality*. However, a re-engineering project is often seen as a good time to change the system requirements. This introduces the full range of forward engineering activities into a re-engineering project. The problem is that two logically separate tasks are being blended into one.

One solution is to first re-engineer a system without changing the functionality and then upgrade the system by changing the functionality of the target system. This means two systems are produced, the first is an intermediate target system that is functionally equivalent to the existing system. The target system is then upgraded to reflect the necessary requirement changes. The problem is the amount of time required to create the final target system.

Another solution is to blend forward and re-engineering together into one process. This will work if reverse engineering can be used to reconstruct a specification of the system. Requirements analysis can be done and the reconstructed specification changed. Then normal forward engineering practices can be used. However, at this time we really are not able to reconstruct a specification from source code for large real world systems. We are better at reconstructing designs. Should a reconstructed design be modified to incorporate requirement changes?

These are some of the issues that must be addressed as research into software re-engineering continues. More knowledge gleaned from re-engineering project is needed. An understanding of different re-engineering approachs is required. The answers to these issues will lead to a fuller understanding of the software re-engineering process.

Appendix A

Summary of Z Notation

Z notation is based on typed set theory and first-order logic.

Schemas

Z provides a construct called a schema, to describe a specification's state and operations. A schema groups together variable declarations with a list of predicates that constrain the variable's possible values. The predicates also state relationships between variables.

There are three types of schema.

Vertical Schema

```
┌─ SchemaName ──────────────────────────────
│ Variable declarations
├──────────────
│ Predicates on variables
└───────────────────────────────────────────
```

Horizontal Schema

Schema Name $\widehat{=}$ [*Variable declarations* | *predicates*]

Axiomatic Schema

```
│ Declaration
├──────────
│ Predicates
```

The axiomatic schema is used to define global functions and constants.

Other notation used with schemas are shown below.

p? : T	States that p of type T is input to a operation denoted by a schema.
p! : T	States that p of type T is output by an operation denoted

	by a schema.
p	Denotes the value of p before an operation begins.
p'	Denotes that the value of p after an operation is complete.
[x / t]	Substitute all occurrences of variable t with variable x.

Definitions and Declarations

$L \mathrel{\widehat{=}} R$	Defines L to be equivalent to R.
v : T	Declares v to be of type T.

Sets

$\varnothing$	The empty set.
$\mathbb{P}\, X$	The powerset of set X. The set of all subsets of X.
$\mathbb{F}\, X$	The set of finite subsetx of X.
$S \subset T$	The set S is a subset of the set T, and S is not equal to T.
$S \subseteq T$	The set S is a subset of the set t or equal to T.
$x \in S$	The item x is an element of the set S.
$x \notin S$	The item x is not an element of the set S.
$S \cup T$	The union of sets S and T.
$S \cap T$	The intersection of sets S and T.

Relations

$x \mapsto y$	(x,y) is an element of a relation.
dom R	The domain of a relation.
ran R	The range of a relation.
$S \lhd R$	Domain restriction. The domain of relation R is restricted such that it equal to set S.
$S \mathbin{\lhd\!\!\!-} R$	Domain subtraction. The domain of relation R is restricted such that it does not contain an element from set S.
$R_1 \oplus R_2$	Overriding, given $R_1, R_2 : X \rightarrow Y$ $\mathrel{\widehat{=}}$ (dom $R_2 \mathrel{\lhd\!\!\!-} R_1$) $\cup R_2$.

Functions

$X \nrightarrow Y$	The set of partial functions from X to Y.
$X \rightarrowtail\!\!\!\!\!\rightarrow Y$	The set of one-to-one partial functions from X to Y. No two elements of X map to the same element of Y.
$f(x)$	The function f given value x.

Bibliography

[Ashok 89] Ashok, V., Ramanathan, J., Sarkar, S., Venugopal, V., "Process Modelling in Software Environments," *ACM SIGSOFT Software Engineering Notes*, (4th International Software Process Workshop) Vol. 14, No. 4, June 1989, pp. 39 - 42.

[Boehm 88] Boehm, Barry W., "A Spiral Model of Software Development and Enhancement," *Computer*, Vol. 21, No. 5, May 1988, pp. 61 - 72.

[Boehm 89] Boehm, Barry W., Bels, Frank, "Applying Process Programming to the Spiral Model," *ACM SIGSOFT Software Engineering Notes*, (4th International Software Process Workshop) Vol. 14, No. 4, June 1989, pp. 46 - 56.

[Byrne 90] Byrne, Eric J., *A Software Re-engineering Process Model Based on Reverse Engineering*, Final Report, USAF-UES Summer Graduate Student Research Program, August 1990.

[Chikofsky 90] Chikofsky, Elliot J., Cross, James H., "Reverse Engineering and Design Recovery: A Taxonomy," *IEEE Software*, Vol. 7, No. 1, January 1990, pp. 13 - 17.

[Curtis 87] Curtis, Bill, Krasner, Herb, Shen, Vincent, Iscoe, Neil, "On Building Software Process Models Under the Lamppost," *International Conference on Software Engineering*, Monterey California, (March 30 - April 2, 1987), pp. 96 - 103.

[Dowson 86] Dowson, Mark, "The Structure of the Software Process," (2nd International Workshop on the Software Process and Software Environments), *ACM SIGSOFT Software Engineering Notes*, Vol. 11, No. 4, August 1986, pp. 6 - 8.

[Gamalel-Din 88] Gamalel-Din, Shehab A., Osterweil, Leon J., "New Perspectives on Software Maintenance Processes," *Conference on Software Maintenance*, Phoenix, Arizona, October 24-27, 1988, pp. 14 - 22.

[Hayes 87] Hayes, Ian (ed.), *Specification Case Studies*, Prentice-Hall, (Englewood Cliffs, New Jersey), 1987.

[Houtz 83] Houtz, Carol A., "Software Improvement Program (SIP): A Treatment for Software Senility," *Proceedings of the 19th Computer Performance Evaluation Users Group*, National Bureau of Standards Special Publication 500-104, October 1983, pp. 92 - 107. (Reprinted in Tutorial on Software Restructuring).

[Iivari 87] Iivari, Juhani, "A Hierarchical Spiral Model for the Software Process," *ACM SIGSOFT Software Engineering Notes*, Vol. 12, No. 1, January 1987, pp. 35 - 37.

[ISPW 88] Tully, Colin (ed.), *4th International Software Process Workshop*, (Moretonhampstead, Devon, UK, May 11-13, 1988), ACM Software Engineering Notes, Vol. 14, No. 4, June 1989.

[Katayama 81] Katayama, Takuya, "HFP: A Hierarchical and Functional Programming Based on Attribute Grammar," *5th International Conference on Software Engineering*, San Diego, California, March 9-12, 1981, pp. 343 - 353.

[Katayama 89] Katayama, Takuya, "A Hierarchical and Functional Approach to Software Process Description," *ACM SIGSOFT Software Engineering Notes*, (4th International Software Process Workshop), Vol. 14, No. 4, June 1989, pp. 87 - 92.

[Kellner 89] Kellner, Mark I., "Representation Formalisms for Software Process Modeling," *ACM SIGSOFT Software Engineering Notes*, (4th International Software Process Workshop), Vol. 14, No. 4, June 1989, pp. 93 - 96.

[Lehman 87] Lehman, M. M., "Process Models, Process Programs, Programming Support," *International Conference on Software Engineering*, Monterey, California, March 30 - April 2, 1987, pp. 14 - 16.

[Mi 90] Mi, Peiwei, Scacchi, Walt, "A Knowledge-Based Environment for Modeling and Simulating Software Engineering Processes," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 2, No. 3, September 1990, pp. 283 - 294.

[OSD/FCSC 83] Federal Conversion Support Center, General Services Administration, *Guidelines for Planning and Implementing a Software Improvement Program (SIP)*, Report No. OSD/FCSC-83/004, May 1983.

[OSD/FSMSC 89] U.S. General Services Administration, Federal Software Management Support Center, *Conversion Management Handbook*, September 1989.

[Osterweil 87] Osterweil, Leon, "Software Processes are Software Too," *International Conference on Software Engineering*, Monterey, California, (March 30 - April 2, 1987), pp. 2 - 13.

[Ramanathan 88] Ramanathan, Jayashree, Sarkar, Soumitra, "Providing Customized Assistance for Software Lifecycle Approahces," *IEEE Transactions on Software Engineering*, Vol. 14, No. 6, June 1988, pp. 749 - 757.

[Sneed 90] Sneed, Harry M., Kaposi, Agnes, "A Study on the Effect of Reengineering upon Software Maintainability," *Conference on Software Maintenance*, San Diego, CA, November 26-29, 1990, pp. 91 - 99.

[Spivey 89] Spivey, J.M., *The Z Notation : A Reference Manual*, Prentice-Hall,(Hertfordshire, England), 1989.

[Sutton 89] Sutton, W. Linwood, "Advanced Models of the Software Process," *ACM SIGSOFT Software Engineering Notes*, (4th International Software Process Workshop), Vol. 14, No. 4, June 1989, pp. 156 - 158.

[Taylor 88] Taylor, Richard N., et al., "Foundations for the Arcadia Environment Architecture," *ACM SIGSOFT Software Engineering Notes*, (Software Engineering Symposium on Practical Development Environments), Vol. 13, No. 5, November 1988, pp. 1 - 13.

[Wileden 86] Wileden, Jack C., "This Is IT: A Meta-Model of the Software Process," *ACM SIGSOFT Software Engineering Notes*, (International Workshop on the Software Process and Software Environments), Vol. 11, No. 4, August 1986, pp. 9 - 11.

[Williams 88] Williams, Lloyd G., "Software Process Modeling: A Behavioral Approach," *International Conference on Software Engineering*, Singapore, (April 11-15, 1988), pp. 174 - 186.

[Yu 90] Yu, Weider D., "A Modeling Approach to Software Cost Estimation," *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 2, February 1990, pp.309 - 314.

FINAL REPORT

PROBABILITY MODELING IN AUTOMATIC TARGET RECOGNITION

1990-91 RESEARCH INITIATION PROGRAM
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Prepared by:	Dr. R. H. Cofer
	(407) 768-8000, ext 8818
Academic Rank:	Associate Professor
Department:	Electrical Engineering
University:	Florida Institute of Technology
	150 West University Boulevard
USAF Researcher:	Ms. Lori Westerkamp

Acknowledgments

I thank the Automated Target Recognition Branch of the Wright Laboratories at the Wright-Patterson Air Force Base and the Air Force Office of Scientific Research for sponsoring this research. Universal Energy Systems should also be mentioned for their efficient administration of this contract.

I would like to thank Ed Zelnio, Lori Westerkamp, and Jim Leonard for allowing me to participate as a member of their research team.

PROBABILITY MODELING IN AUTOMATIC TARGET RECOGNITION

Dr. R. H. Cofer

1.0 BACKGROUND

This is the latest in a series[7-11] of research efforts developing the theory for high performance ATR systems. It is based on a theoretically provable, optimally accurate approach to ATR -- that of Bayesian model matching. The optimality of the Bayesian approach lies in its theoretical ability to accumulate and rigorously fuse all-source evidence: incoming multi-sensor imagery; target, scene, and sensor models and collateral information.

Florida Institute of Technology, FIT, is primarily working on the advanced theoretical issues shown in Figure 1. The Target Recognition Group, AARA, of Wright Laboratories, is studying the practical implementation issues.

The 1989 SFRP period successfully laid the keystone block of Bayesian probability decomposition.[7] This continues to permit attack of individual subproblems with assurance that the results can be fitted together. Chief of these subproblems is the determination of joint probability density functions of targets since these functions permit the most accurate utilization of image evidence for automatic target recognition, ATR.

Reference 11 documents very successful results achieved under the 1989-90 RIP effort toward solving the LADAR target probability characterization problem.

The 1990 SFRP program[10] developed considerable insight into robustness of the Bayesian approach to modeling errors. It also developed insights into the IR ATR problem including the concepts of underlying and local homogeneity of target temperatures.

2.0 THE CURRENT PROBLEM BEING RESEARCHED

High performance recognition of targets in IR imagery continues to be a strong goal of AARA and thus the subject of this current research.

The key to higher performance recognition is the detailed understanding of just how the target can appear in the image.[7] - necessitating careful modeling of the scene formation and sensing processes.

An IR image is formed as a result of imaging the temperatures, T's, of a scene, Figure 2. If the target in the scene were to have constant temperature throughout, then its visual detail within the IR image would disappear and thus become impossible to recognize.

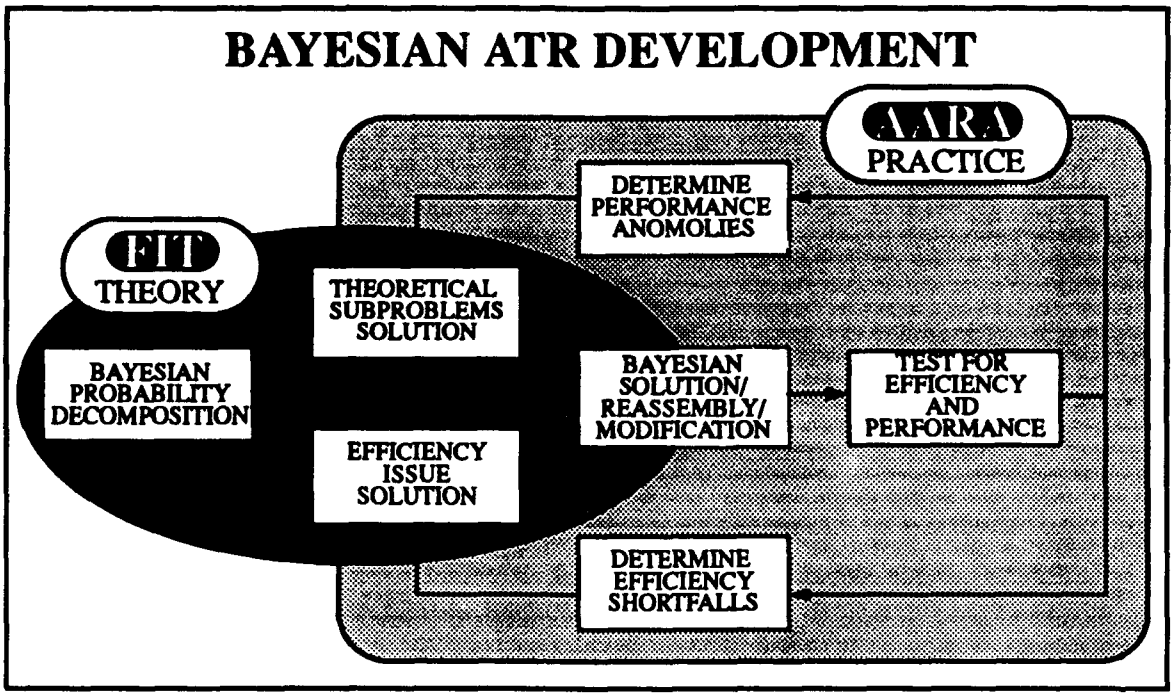


Figure 1. Long Term Research Strategy

Fortunately, flows of heat to and from targets are continuously varying due to the diurnal thermal cycle and the random vagrancies of wind, cloud cover, air temperature, rain and the temperature of the target's surroundings. These external heat flows set up heat flows within the target in accordance with the conduction equation

$$\frac{dT}{dt} = \frac{\partial^2 T}{\partial z^2} + \sum_i \frac{\partial q_i}{\partial z} \qquad \text{(EQ 1)}$$

The resulting heat flows internal to the target cause corresponding variations of temperature within the target. It is these variations of target temperature which provide recognizable target detail in IR imagery.

Unfortunately, much of the flow of heat to and from the target is random, causing random variations in target temperature and thus random variation of the visual detail of targets within IR imagery. This results in a problem for ATR. Since one can not know just how the target will appear in the IR image, how can it be recognized with any certainty? The answer, rather reasonably, is that the target should be recognized as residing in any IR image that itself has high probability of arising from the target.

The problem for current research reduces now to "How to obtain the critical probability function of the target's surface temperatures?" Given the years of IR target modeling to date, this is surprisingly still an open question. Nevertheless, study shows that previous IR

image modeling efforts have concentrated on finding a target's IR appearance given full knowledge of its thermal history. Since one will never have 'full knowledge of the thermal history' of enemy targets on the battlefield, the utility of these prior efforts for ATR is problematical.

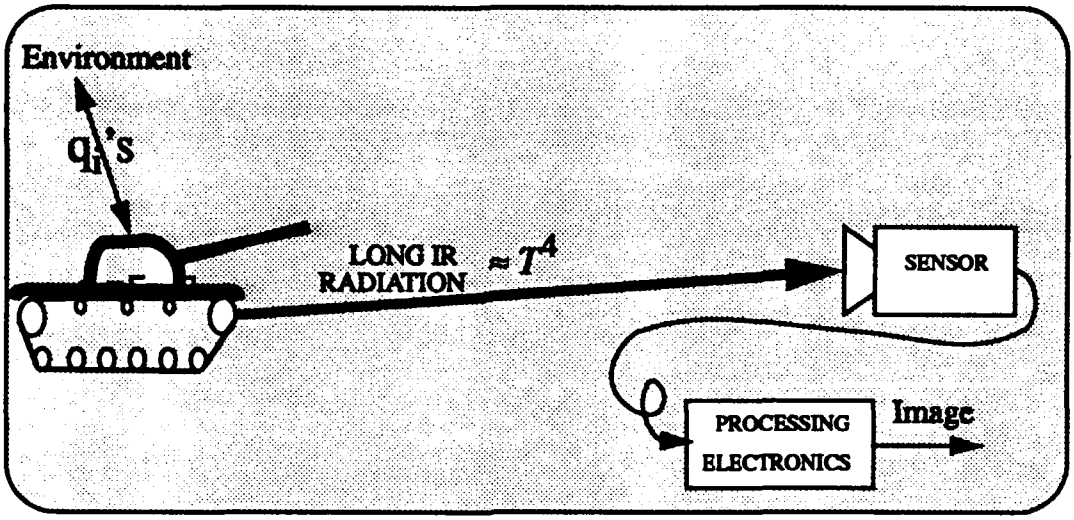


Figure 2. The IR Sensing Process

3.0 THE RESEARCH TASKING

The following tasks for the current research effort were originally proposed in accordance with guidance from AARA:

- Task 1 - Determination of Underlying Temperatures of IR Targets

  Description: Use all available means to move toward meaningful theoretical modeling characterization of the major underlying temperatures of IR targets.

- Task 2 - Characterization of Near BIOT Conditioning

  Description: Theoretically consider the development of the probabilities of the surface temperature homogeneity of target parts under near BIOT conditioning.

- Task 3 - Develop LADAR Target Detection and Recognition Bounds

  Description: Develop ROC and recognition confusion matrix bounds on LADAR target detection and recognition, respectively.

As the current research effort began, AARA had already begun to develop code for ROC and recognition bounds for LADAR targets as a result of earlier FIT theory. AARA thus determined that FIT would act as a consultant on demand against the scope of task 3, while placing the maximum effort against tasks 1 and 2.

As originally conceived, tasks 1 and 2 contained both experimental and theoretical considerations. Since there was no reasonable avenue for the Government to provide the neces-

sary equipment for timely experimental investigations, it was jointly agreed that the current effort at FIT should be totally theoretical in nature.

A multi-step theoretical approach was then undertaken which has produced much greater success than reasonably to be expected - given the distributed thermal memory of an IR target. It has become possible to develop techniques for determining the joint probability density function of IR target regions given only the statistical characterization of their thermal environmental drivers. There is no recourse to Gaussian approximations or restriction to first and second moments. Determination of these probability functions for targets will find wide use in the field of IR ATR.

4.0 TECHNICAL RESULTS

The results of this research are sets of mathematical results gathered into appendices: processes which can be used in a variety of situations to gain insight and derive probabilities of target presence over a wide range of IR sensing and environmental situations.

The first set of results, Appendix A, transforms the thermal problem in solids, such as metal, to a form more easily manipulated: that of resistor-capacitor passive electrical circuits - either in distributed or lumped form. This action, for electrical engineers at least, allows more intuitive consideration.

The second set of results, Appendix B, uses the tools of modern signal analysis to determine target thermal responses to planar environmental drivers. Distributed system responses to sinusoidal, step, and impulse inputs are found. These results enable immediate use of Fourier and convolution methodologies to develop exact target thermal responses to arbitrarily complex input planar drivers without recourse to finite element approximations. Quantitative expressions for the speed of thermal propagation and attenuation of the traveling signal in target solids are given. It is seen that the thermal wave dies out quite rapidly with an attenuation of 55 db per wavelength. The wavelength of the thermal wave is seen to be a function of the specific heat of the solid and the frequency of the externally applied thermal driver. Examples are given which quantify target potentials for signal propagation within targets over frequencies on the order of the daily diurnal cycle down to brief wind gusting.

The third set of results, Appendix C, considers the thermal accuracy of lumped constant modeling of solids, Appendix E, with respect to the exact results obtained in Appendix B. Here it is seen that a lumped stage per 2.6 cm of target distance at brief wind gusting frequencies and per 1/2 meter at diurnal frequencies is more than sufficient to maintain reasonable signal accuracy. These results are important for the remaining appendices, which utilize lumped rather than distributed modeling.

The forth set of results, Appendix D, is concerned with extension to spatially distributed drivers over the target body, i.e. a thermal bath. This is achieved by spatial superposition and voltages (temperatures) of the capacitor analogs as state variables. This allows state equation representation of arbitrarily large target solids while removing restrictions on tar-

get homogeneity. Specific attention is given to developing a straight-forward and mechanical approach for reducing arbitrarily complex target solid models to the state equation formulation.

The fifth set of results, Appendix E, is devoted to obtaining the impulse response of the target solid as represented by the state equations and while responding to a spatially and temporarily diverse set of thermal drivers. The impulse response is found to be in matrix form and is used to develop the matrix convolution equation capable of giving any joint set of thermal outputs from the target solid in response to arbitrary thermal baths. This is a key result, as analytic expressions of joint sets of outputs is required by the next two appendices.

The sixth set of results, Appendix F, provides the key results of the current study: the derivation of means of determining the joint probability density functions of target regions in IR imagery - given only statistical knowledge of the thermal environmental drivers. This is accomplished for all three cases: equal numbers of thermal drivers and IR target pixels, more thermal drivers than IR target pixels, and less numbers of thermal drivers than IR target pixels. The question, "Does one needs to know the full statistical behavior of the environmental driver set or is some law of large numbers at work alleviating the need for full statistical knowledge?" is answered for the first subcase. For equal numbers of thermal drivers and IR target pixels, Appendix F shows one indeed needs the full statistical knowledge. Appendix F then shows for more thermal drivers than IR target, there is a "law of projection" at work which includes the law of large numbers as a special case. Given the "law of projection," it is shown that the complexity of the IR target pixels joint probability density function can be greatly reduced from that of the environmental drivers. Appendix F also shows how to solve the subcase of more IR target pixels than environmental drivers. Finally this appendix shows extensions to the theory also useful in the design of ATR systems including incorporation of low sensor resolution, and characterizing thermal spatial gradients.

The seventh and final set of results, Appendix G, considers the effects of nonlinear coupling between the target and its surrounding environment on the above techniques for deriving joint probability density functions. It gives means for deriving the IR target joint probability density function under nonlinear environmental coupling. These new techniques will require additional mathematical steps but each step will be shorter and potentially less complex.

5.0 CONCLUSIONS

The current research, reported here, represents a significant step toward solving IR ATR problems. Specifically, it overcomes the problem to date in IR model matching that the complete thermal environment history is needed to derive a synthetic view of the target model for ATR. Given only the statistics of the thermal environment drivers acting on the target, techniques are shown that will mathematically give the joint probability of the target IR region as imaged and as conditioned on specific target modeling and indexing.

Additionally, the mathematical techniques which are presented are theoretically capable of solving nonlinear couplings to the environment. All of the resulting techniques avoid need to develop sample statistics via Monti-Carlo simulation. This is fortunate since Monti-Carlo methods are inappropriate due to the high dimensionality of the problem.

6.0 RECOMMENDATIONS

Recommendations based on the current research lie in three areas: extension of the theory, application of the theory, and ATR systems implementation. Other strongly related recommendations will be found in prior reports of this series.[7-11]

Recommended extensions to the currently developed theory include:

- Developing statistics of the thermal environment. Specifically needed are the spatial distribution and temporal independencies of thermal drivers.
- Further amplification of the concept of the "projection law" for joint probability distribution functions. There is vanishingly little discussion in the statistical literature of the degree of complexity that joint probability density functions can achieve in nature and how this complexity is compressed out in moving to lower order signal spaces, yet it is of first order importance to the ATR problem.
- Further study of the entire concept of linearizing the problem under nonlinear environmental couplings. Under what conditions does it make sense, how can the concept be expanded, etc.?
- Further mathematics should be developed which will lead to algorithmic methods for handling nonlinear environmental couplings with less effort and more clarity.
- Use of modern circuit synthesis techniques that provide accuracy with fewer "circuit" components.
- Conversion of all results to the distributed case by passing to the infinitesimal limit.

Recommended applications of the theory include:

- Using the joint probability density function of IR target regions as features within various formulations of ATR systems, including final evidence accumulation.
- Using the joint probability density functions of IR target regions to develop bounds on recognition accuracy achievable by any ATR system.
- Using the joint probability density functions of IR target regions to effect target-decoy discriminations.
- Continuing research toward a truly optimal IR ATR system based on Bayesian minimum error of recognition concepts by researching the IR statistical characteristics of background.
- Merging the results of the current research with those of past efforts in this series to better allow pixelwise sensor fusion.
- Use of joint probability density functions to aid in detecting target index mismatch.

- Use of Biot inspired macro modeling to determine joint statistics of the underlying joint temperature distributions of major target components.

Recommendations for further research into system implementation issues include:

- Determination of the crossover point, if any, at which the complexity of learning the joint probability density function of target regions becomes less that the calculation from first principles developed here.
- Determination of the likely complexities of the joint probability density function of target regions and their impact on ATR algorithmic accuracy vs. storage and training requirements.
- Study of efficiency gains due to approximation of the probability density function of target regions at various levels of IR resolution.

Appendix A
ELECTRICAL ANALOGS OF THERMAL CONDUCTION

A1.0 INTRODUCTION

The experimental law for heat conduction transfer through solids is due to Fourier. The law states that the amount of heat flowing through a surface per unit time is proportional to the temperature gradient normal to the surface

$$\frac{\partial Q}{\partial t} = -k\Delta A\frac{d\tau}{dx} \tag{A1}$$

where Q is the quantity of heat in joules, t is time in seconds, k is conductivity in watts/m-C, τ is the temperature in degrees C and the geometry is shown in Figure A1.

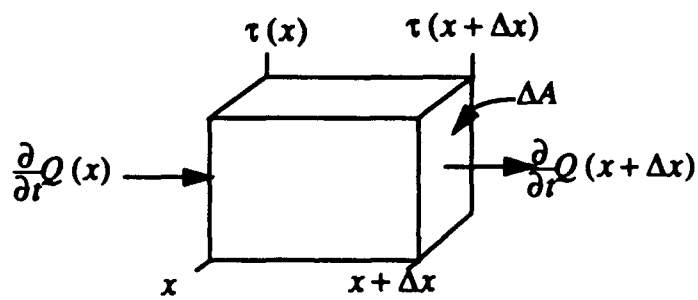


Figure A1 An Infinitesimal Volume Of A Solid

A2.0 CONVERSION TO AN ELECTRICAL CURRENT

In order to convert the thermal conduction problem to one of electrical analogies, it is convenient to express heat as a vector rate of flow per unit area, q

$$q = \lim_{\Delta A \to 0} \frac{1}{(\Delta A)^2}\frac{\partial Q}{\partial t}\Delta A \tag{A2}$$

This is a flux density corresponding to electrical current density.

A3.0 CONVERSION TO A RESISTOR

If a plane wave of heat transfer is assumed, so only one component of the vector, q, is needed, then Eq.(A1) becomes

$$q = -k\frac{\partial \tau}{\partial x} \tag{A3}$$

or when rewritten

$$\frac{1}{k} = \frac{\frac{\partial \tau}{\partial x}}{-q} \tag{A4}$$

one can make the conversion

$$R = \frac{v}{i} \tag{A5}$$

to give the electrical analog of a resistor, Figure A2.

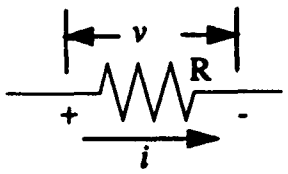


Figure A2 Notation Relating To A Resistor

A4.0 CONVERSION TO A CAPACITOR

Consider Figure A1. The heat flowing in from the left is

$$\frac{\partial}{\partial t}Q(x) = q(x)\,\Delta A \tag{A6}$$

and the heat flowing out from the right is

$$\frac{\partial}{\partial t}Q(x+\Delta x) = q(x+\Delta x)\,\Delta A \tag{A7}$$

The difference between these two values represents the amount of heat, $S\rho\Delta V$,being removed from the volume, ΔV. S is termed the specific heat per unit mass, i.e. a constant giving the change in heat stored within a given mass to the change in temperature, and ρ is the density in kilograms per cubic meter of the mass

$$\frac{\partial}{\partial t}Q(x+\Delta x) - \frac{\partial}{\partial t}Q(x) = q(x+\Delta x)\,\Delta A - q(x)\,\Delta A = -S\rho\Delta V \tag{A8}$$

In passing to the limit

$$\frac{\partial q}{\partial x} = -S\rho\frac{\partial \tau}{\partial t} \tag{A9}$$

and upon rewriting

$$i = C\dot{v} \tag{A10}$$

one see the relation to the electrical analog of a capacitor, Figure A3.

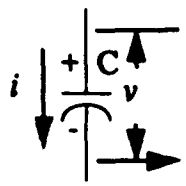


Figure A3 Notation Pertaining To A Capacitor

A5.0 CONVERSION TO CIRCUIT ELEMENTS

Eq.'s (A5) and (A10) taken together give the low pass RC filter, Figure A4, which is

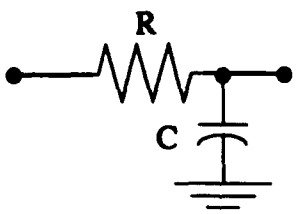


Figure A4 The RC Filter Analog

equivalent to the thermal diagram of Figure A1.

To develop more complex circuits, divide the solid into subvolumes placing a grounded capacitor in each. Then place a resistor between capacitors of adjoining subvolumes.

As an example, consider planar heat flow through a plate. The equivalent circuit is shown in Figure A5.

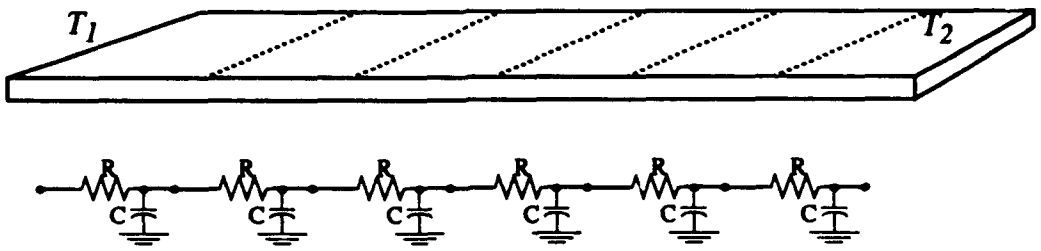


Figure A5 Example Conversion To The RC Filter Analog

Appendix B
SIGNAL ANALYSIS OF THERMAL CONDUCTION

B1.0 INTRODUCTION

Appendix A gave the electrical analog of thermal conduction in solids. As a first step to the understanding of probabilistic heat flows, this appendix uses modern signal analysis techniques to derive further understanding of the conducting heat flow signal itself.

B2.0 RESPONSE TO STEADY-STATE ALTERNATING TEMPERATURES

The target in the field is subjected to various alternating environmental drivers - the chief of which is the diurnal thermal cycle. As a first step of toward understanding, this section considers the response of a flat target plate to an alternating sine wave of temperature, T_1, in Figure A5. Since temperature can not go negative, all signals of this section are with respect to some average temperature.

Combining Eq.'s (A5) and (A10) gives

$$\frac{\partial^2 v}{\partial x^2} = RC\frac{\partial v}{\partial t} \tag{B1}$$

As usual, let the instantaneous values of steady state alternating voltage and current be

$$v = Re(Ve^{j\omega t}) = V\sin\omega t \tag{B2}$$

$$i = Re(Ie^{j\omega t}) = V\cos\omega t \tag{B3}$$

which when substituted into Eq. (B1) gives at any angular frequency, ω, the harmonic equation

$$\frac{d^2 v}{dx^2} = j\omega RCv = \Upsilon^2 v \tag{B4}$$

where Υ is called the propagation constant.

In terms of heat quantities Υ is

$$\Upsilon = \sqrt{\frac{j\omega S\rho}{k}} = \sqrt{\frac{\omega}{2D}}(1+j) \tag{B5}$$

where D is $\frac{S\rho}{k}$, the thermal diffusivity.

The well know solution[1] to the harmonic equation is

$$v = Ae^{-\Upsilon x} + Be^{\Upsilon x} \tag{B6}$$

Now, solving for i from Eq. (A4)

$$\begin{aligned} i &= \frac{\frac{\partial v}{\partial x}}{R} \\ &= \frac{1}{R}(A\Upsilon e^{-\Upsilon x} - B\Upsilon e^{\Upsilon x}) \\ &= \frac{1}{\sqrt{\frac{R}{j\omega C}}}(Ae^{-\Upsilon x} - Be^{\Upsilon x}) \\ &= \frac{1}{Z_0}(Ae^{-\Upsilon x} - Be^{\Upsilon x}) \end{aligned} \tag{B7}$$

where Z_0 is the impedance of the solid. In thermal terms

$$Z_0 = \sqrt{\frac{1}{j\omega kS\rho}} = \sqrt{\frac{1}{2\omega kS\rho}}(1-j) \tag{B8}$$

which shows that the flow of heat is always 45 degrees ahead of the temperature.

Assuming the solid plate is of infinite length, then B must be zero in Eq. (B7), otherwise the $e^{\Upsilon x}$ term would grow to infinity as $x \to \infty$. The $e^{\Upsilon x}$ term comes into play for finite plate length only in terms of signal reflections. Signal reflections will be of little interest in the thermal case for reasons seen in Subsection B2.2.

Since B must be zero valued

1. An alternate solution form in terms of hyperbolic functions can be given

$$v = C\cosh\Upsilon x + D\sinh\Upsilon x$$

This alternate formulation tends to destroy the meaning for the thermal problem, and so will be avoided in this work.

$$v = Ae^{-\Upsilon x}$$
$$i = \frac{1}{Z_0} A\Upsilon e^{-\Upsilon x} \tag{B9}$$

and of course

$$Z_0 = \frac{v}{i} \tag{B10}$$

Let Υ be represented as

$$\Upsilon = \alpha + j\beta \tag{B11}$$

where α is termed the attenuation constant and β is the phase constant.

In electrical terms

$$\alpha = \beta = \sqrt{\frac{\omega RC}{2}} \tag{B12}$$

In thermal terms

$$\alpha = \beta = \sqrt{\frac{\omega S\rho}{2k}} = \sqrt{\frac{\omega}{2D_t}} \tag{B13}$$

In ordinary circuit analysis, the time factor is frequently neglected (as above). Since in a solid there are both time and space factors, it is best to include both explicitly. So from Eq.'s (B9) and (B11)

$$ve^{j\omega t} = Ae^{-\alpha x}e^{j(\omega t - \beta x)}$$
$$ie^{j\omega t} = \frac{A}{Z_0}(e^{-\alpha x}e^{j(\omega t - \beta x)}) \tag{B14}$$

or when expressed in real terms, as a function of x and t

$$v(x, t) = Ae^{-\alpha x}\cos(\omega t - \beta x)$$
$$i(x, t) = \frac{A}{Z_0}e^{-\alpha x}\cos(\omega t - \beta x - \theta) \tag{B15}$$

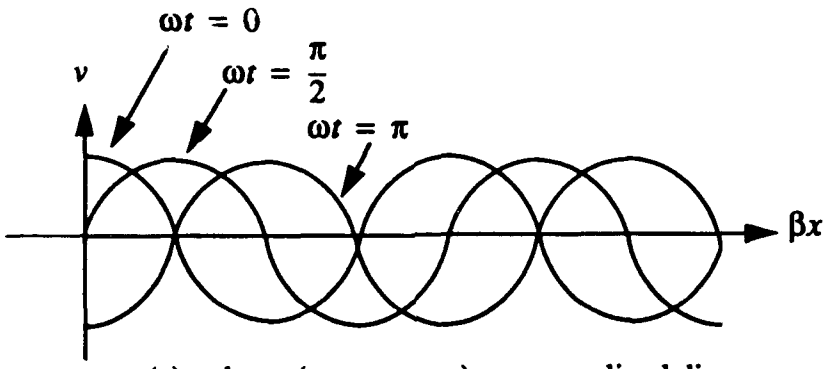
where

$$Z_0 = |Z_0|e^{j\theta} \tag{B16}$$

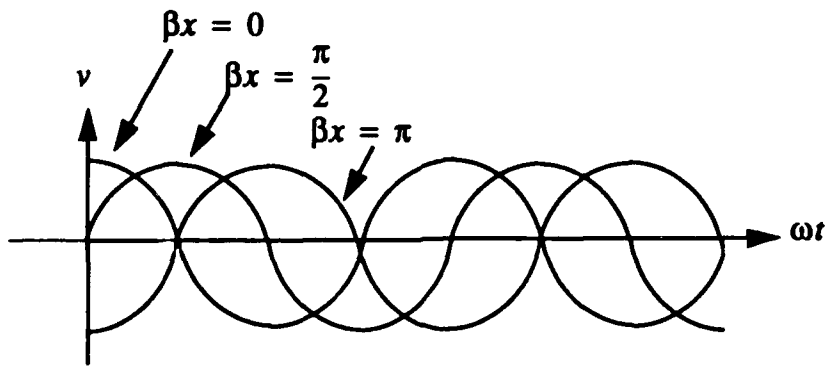
B2.1 Propagation Of The Thermal Wave

To see the meaning of the time/distance relationship of the signal in the solid, concentrate on the term, $e^{j(\omega t - \beta x)}$ of Eq. (B15). Figure B1a shows a plot of the voltage (temperature) across the normalized length, βx, of the solid for three different normalized times, $\omega t = 0$, $\omega t = \frac{\pi}{2}$ and $\omega t = \pi$.

Figure B1b shows a plot of the voltage (temperature) as a function of normalized time, ωt, for three different normalized positions on the solid. Comparison of the two sets of plots reveals that they are the same in shape[1]. Figure B1 is thus a picture of a traveling wave of temperature launched across the solid. The wavelength in time is $\frac{2\pi}{\omega}$ and the wavelength



(a) voltage (temperature) vs. normalized distance



(b) voltage (temperature) vs. normalized time

Figure B1 Spatial and Temporal Snapshots Of The Traveling Wave

in space is $\frac{2\pi}{\beta}$. The space wavelength in electrical terms is

1. Actually this is a partial result of the use of an even function, the sinusoid. In general the space and time plots are mirror images. This will be better seen when dealing with transients in following sections.

$$\lambda = \sqrt{\frac{8\pi^2}{\omega RC}} \tag{B17}$$

In thermal terms

$$\lambda = \sqrt{\frac{8\pi^2 k}{\omega S\rho}}$$

$$= \sqrt{\frac{8\pi^2 D}{\omega}} \tag{B18}$$

The most important characteristic of this equation is that it shows the wave length of thermal propagation in solids to be a direct function of the driving (time) frequency. The higher the driving frequency the faster the propagation and the shorter the wavelength.

The phase velocity, v_p, is the velocity at which the thermal wave is moving. It can be found from setting $\omega t - \beta x$ in Eq. (B15) to zero giving

$$v_p = \frac{\partial v}{\partial t} = \frac{\omega}{\beta} \tag{B19}$$

In electrical terms

$$v_p = \sqrt{\frac{2\omega}{RC}} \tag{B20}$$

and in thermal terms

$$v_p = \sqrt{\frac{2\omega k}{S\rho}} = \sqrt{2\pi D} \tag{B21}$$

B2.2 Attenuation Of The Thermal Wave

In Eq. (B15), the envelope term of the thermal wave is $e^{-\alpha x}$. To see how much the thermal wave is reduced in amplitude as it travels across the solid, consider that at x the voltage is

$$v_x = ce^{-\alpha x} \tag{B22}$$

and at $x + l$, it is

$$v_{x+l} = ce^{-\alpha(x+l)} \tag{B23}$$

The ratio of these two envelopes is

$$\left|\frac{v_{x+l}}{v_l}\right| = e^{-\alpha l} \tag{B24}$$

for any x. This gives an attenuation of

$$\alpha l = \log_e \frac{v_{x+l}}{v_x} \tag{B25}$$

in nepers per length l. The equivalent attenuation of the traveling wave in db is

$$20\log_{10} \frac{v_{x+l}}{v_x} = (20\log_{10} e)\,\alpha l \tag{B26}$$

Taking l to be one wavelength gives

$$\alpha l = \sqrt{\frac{\omega RC}{2}}\sqrt{\frac{8\pi^2}{\omega RC}} = 2\pi \tag{B27}$$

which is a *very* large attenuation - 2π nepers or 55 db per wavelength. Thus in just one wavelength the absolute attenuation is a factor of 526. Even at one half wavelength is absolute attenuation is quite high, a factor of 25. As shown in Eq. (B18), the absolute distances for these levels of attenuation depends on the driving (time) frequency and the thermal diffusivity for the solid. Figure B2 gives representative values for one wavelength of distance.

Type Solid	Time Wavelength			
	1 Day	1 Hour	1 Minute	1 Second
Copper	10.4 M	2.1 M	27 cm	3.5 cm
Steel	3.3 M	67 cm	8.6 cm	1.1 cm
Concrete	1 M	21 cm	2.7 cm	0.35 cm

Figure B2 One Wavelength In Various Solids And At Various Driving Frequencies

Appendix C
ACCURACY CONSIDERATIONS IN LUMPED CONSTANT MODELING

C1.0 INTRODUCTION

Appendix A developed electrical analogs to the thermal conduction process that can be utilized in the distributed or discrete (lumped constant) case. Appendix B developed signal responses to a variety of environmental drivers based on the distributed analog model. These responses are therefore exactly those of the thermal conduction processes since no approximations have been made. On the other hand, Appendix D through Appendix G are based on the discrete (lumped constant) form of the analog model. This raises accuracy considerations. The lumped constant model must be made fine grained enough to prevent significant error. This appendix develops the necessary theory to make direct comparisons against the exact results of Appendix B.

C2.0 THE APPROACH

A lumped-constant analog model for a planar traveling wave in the solid of Figure A5 is seen to be an iterative structure, i.e. a cascade of identical stages. The solid has the same characteristic impedance, given by Eq. (B8) at every point along its length. The iterative analog in Figure A5 also has a constant repetitive characteristic impedance at the output of each of its stages.

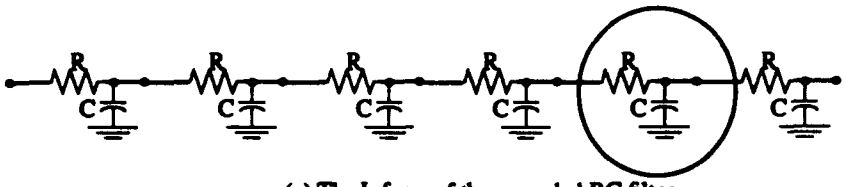
As the number of stages of the analog rises per unit length of the modeled solid, the characteristic impedance and other parameters of the analog converge to those of the solid itself. Thus a reasonable goal is to determine the number of iterative stages required per unit length to achieve a desired fidelity to the actual parameters of the solid itself.

C3.0 PARAMETERS OF THE ITERATED STRUCTURE

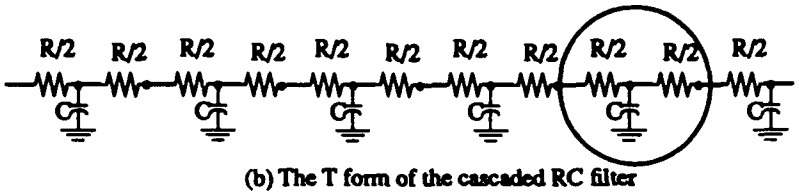
In order to determine the values of the parameters of the iterated analog structure of Figure A5, it is easier to pass to the T form, Figure C1, of the RC filter section prior to performing calculations. R can be taken as the resistance per meter of the solid divided by the number of iterative stages per meter. C can similarly be taken as the capacitance per meter of the solid divided by the number of iterative stages per meter.

C3.1 The Iterative Characteristic Constant

When the T filter is properly terminated, its input impedance, Z_0, is equal to its load impedance. Solving for this case



(a) The L form of the cascaded RC filter



(b) The T form of the cascaded RC filter

Figure C1 Equivalence Of The L And T Cascade

$$Z'_0 = \sqrt{j\omega RC}\sqrt{1 + \frac{R}{4j\omega C}} \tag{C1}$$

From Appendix B, the characteristic impedance, Z_0, of the actual solid is found to be

$$Z_0 = \sqrt{j\omega RC} \tag{C2}$$

As expected, Z'_0 converges to Z_0 as $\omega \rightarrow \infty$.

C3.2 The Iterative Propagation Constant

Υ', the propagation constant of each interative stage, can with sufficient manipulation be found as

$$\Upsilon' = \log_e(1 + \frac{R}{2j\omega C} + \frac{Z'_0}{j\omega C}) \tag{C3}$$

C4.0 A FIRST ORDER ACCURACY ANALYSIS

The thermal wave attenuates by 55 db in one wavelength in a solid, Appendix B, and becomes insignificant after that point. A sufficient number of discrete lumped constant stages should then be used in the first wavelength in the electrical analog to insure good fidelity to attenuation and phase of the true traveling thermal wave of the actual solid. A first order analysis, carried out here, indicates that relatively few stages per unit length are sufficient.

Z'_0 can be converted to the form

$$Z'_0 = Z_0\sqrt{1+\frac{(\Upsilon\Delta x)^2}{4}} \tag{C4}$$

where Υ is the actual propagation constant for the solid and Δx is the number of meters per iterative analog T filter stage. Since Z'_0 should be close in value to Z_0, $\frac{(\Upsilon\Delta x)^2}{4}$ should be selected to be much less in value than 1, or

$$\Delta x \ll \left|\frac{2}{\Upsilon}\right| = 2\sqrt{\frac{D}{\omega}} \tag{C5}$$

Assuming this is so, then expanding the square root in Eq. (C4) by the binomial expansion, and retaining the first two significant terms is sufficient

$$Z'_0 = Z_0(1+\frac{(\Upsilon\Delta x)^2}{8}) \tag{C6}$$

Substituting Eq. (C6) into Eq. (C3), using the significant terms of a series expansion for the natural logarithm, and after considerable further manipulation

$$\Upsilon' = \Upsilon\Delta x - \frac{1}{24}(\Upsilon\Delta x)^3 - \frac{1}{4}(\Upsilon\Delta x)^4 + \ldots \tag{C7}$$

which reduces with sufficient accuracy to

$$\Upsilon' = (\Upsilon\Delta x)\,(1-\frac{1}{24}(\Upsilon\Delta x)^2+\ldots) \tag{C8}$$

Since Υ' is the propagation constant of an interative stage, let Υ'' be the propagation constant for n sections of the iterative cascade of stages

$$\Upsilon'' = n\Upsilon' \tag{C9}$$

and so

$$\Upsilon'' = \Upsilon n\Delta x\,(1-\frac{(\Upsilon\Delta x)^2}{24}+\ldots) \tag{C10}$$

Since from Eq. (B12)

$$\Upsilon = \alpha + j\beta = \sqrt{\frac{\omega RC}{2}}\,(1+j) \tag{C11}$$

then

$$\alpha'' = n\Delta x\alpha Re((1+j)\left(1-\frac{(\Delta x)^2\alpha^2(1+j)^2}{24}+\ldots\right)) \tag{C12}$$

$$\beta'' = n\Delta x\alpha Im((1+j)\left(1-\frac{(\Delta x)^2\alpha^2(1+j)^2}{24}+\ldots\right)) \tag{C13}$$

This reduces to

$$\alpha'' = n\Delta x\alpha(1+\frac{(\Delta x\alpha)^2}{12}) \tag{C14}$$

and

$$\beta'' = n\Delta x\beta(1-\frac{(\Delta x\beta)^2}{12}) \tag{C15}$$

Error in attenuation over the n sections is best expressed as a ratio

$$\frac{\alpha''}{n\Delta x\alpha} = (1+\frac{(\Delta x\alpha)^2}{12}) \tag{C16}$$

and error in radian phase over the n sections is best expressed by subtraction

$$\beta'' - n\Delta x\beta = -\frac{n(\Delta x\beta)^3}{12} \tag{C17}$$

C4.1 Interpretation

Reflection on Eq.'s (C4), (C16) and (C17) will show that it is harder to maintain reasonable phase error than reasonable attenuation or characteristic impedance error. Since the thermally traveling wave will have decayed to -55db over a single wavelength, there is little need to consider the phase error over longer distances. A quite conservative allowable phase error is 36 degrees over this single wavelength, thus

$$\frac{n(\Delta x\beta)^3}{12} = 36\frac{2\pi}{360} \tag{C18}$$

Setting the number of filter stages per wavelength

$$\lambda = n\Delta x = \sqrt{\frac{4\pi D}{f}} \tag{C19}$$

Eq. (C18) reduces to

$$\sqrt{\frac{4\pi D}{f}}(\Delta x^2)\beta^3 = 7.54 \tag{C20}$$

Since

$$\beta = \sqrt{\frac{\pi f}{D}} \tag{C21}$$

Eq. (C20) further reduces to

$$\Delta x = 0.618\sqrt{\frac{D}{f}} \tag{C22}$$

For a target of mild steel, then a conservative lower length per iterative stage is

$$\Delta x = \frac{0.002}{\sqrt{f}} \tag{C23}$$

C4.2 Conclusions

If the highest frequency environmental driver of interest is on the order of one cycle per three minutes, then for mild steel it would be appropriate to insert an iterative stage no more often than one per 2.6 cm. Were the highest frequency of interest on the order of one cycle per day, for mild steel it would be appropriate to insert an iterative stage no more often than one per half meter. These figures are quite conservative. Further analysis would most likely double or triple the minimum required spacings.

Appendix D
STATE REPRESENTATION OF THE HEAT CONDUCTION NETWORK

D1.0 INTRODUCTION

Appendix B determined the thermal traveling wave responses of an infinite length solid to a single planar deterministic thermal driver. This appendix is devoted to a formulation which is to be used in Appendix E to allow determination of the target's response to an arbitrarily large number of arbitrarily placed and complex deterministic thermal drivers.

This appendix also provides a straightforward and mechanical method of converting the RC electrical analogs of target thermal conduction to the new formulation.

D2.0 THEORY

The general form of the state equations of a continuous deterministic system can be given as:

$$x(t) = f(x(t_0), v(t_0, t)) \tag{D1}$$

$$y(t) = g(x(t_0), v(t_0, t)) \tag{D2}$$

In those cases where this continuous deterministic system can be given as a set of linear ordinary differential equations, i.e. thermal conduction, the state equations become:

$$\dot{x}(t) = A(t)x(t) + B(t)v(t) \tag{D3}$$

$$y(t) = C(t)x(t) + D(t)v(t) \tag{D4}$$

Finally when the system is time-invariant, i.e. has constant thermal properties, the state equations can be given as:

$$\dot{x}(t) = Ax(t) + Bv(t) \tag{D5}$$

$$y(t) = Cx(t) + Dv(t) \tag{D6}$$

as illustrated in Figure D1.

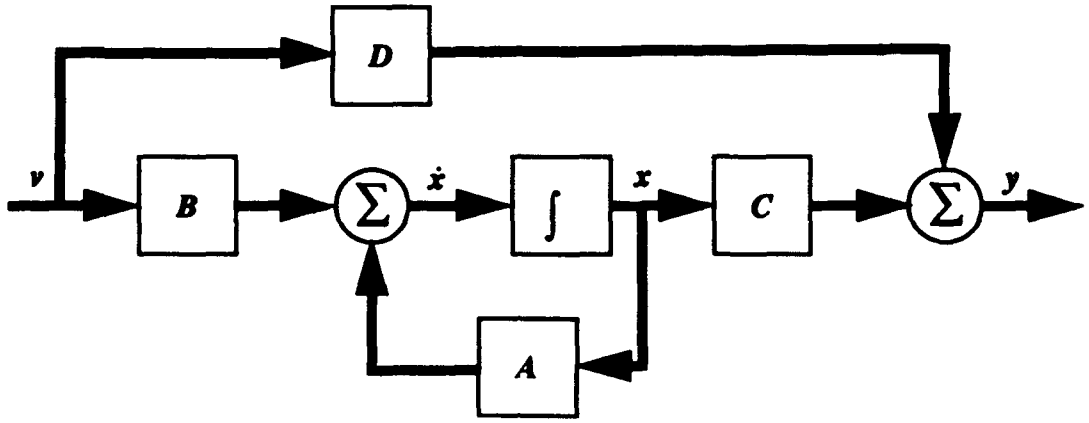


Figure D1 The General State Representation

D3.0 REDUCTION OF HEAT CONDUCTION NETWORKS TO STATE MATRIX REPRESENTATIONS

As discussed in Appendix A, the heat conduction network is equivalent to an electrical resistor capacitor network. Without loss of generality, this resistor capacitor network can be represented as shown in Figure D2

Let

$$x(t) = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \qquad v(t) = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_l \\ i_{l+1} \\ i_{l+2} \\ \cdot \\ \cdot \\ i_m \end{bmatrix} \tag{D7}$$

If all voltage and current sources are set to zero, i.e. respectively short-circuited and open-circuited, and a unit voltage is supplied in series with the j'th capacitor at time t_0, then let the current at t_0+ in capacitor i be noted as $i_{cij}(t_0+)$. Since capacitors initially appear as short-circuits, $i_{cij}(t_0+)$ can be computed as shown in Figure D3. Since $i_{cij}(t_0+)$ is a current through the i'th capacitor, then $c_{cij}(t_0+) = -C_i\dot{x}_i(t_0+)$.

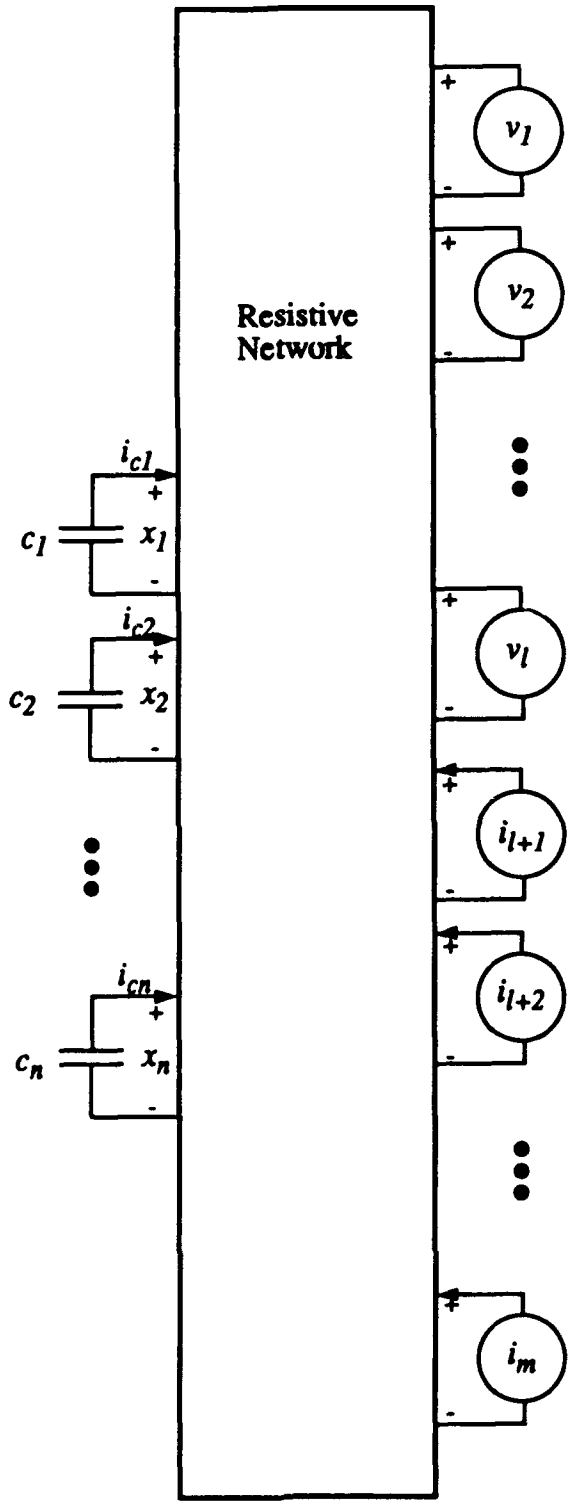


Figure D2 Isolation Of The Purely Resistive Network

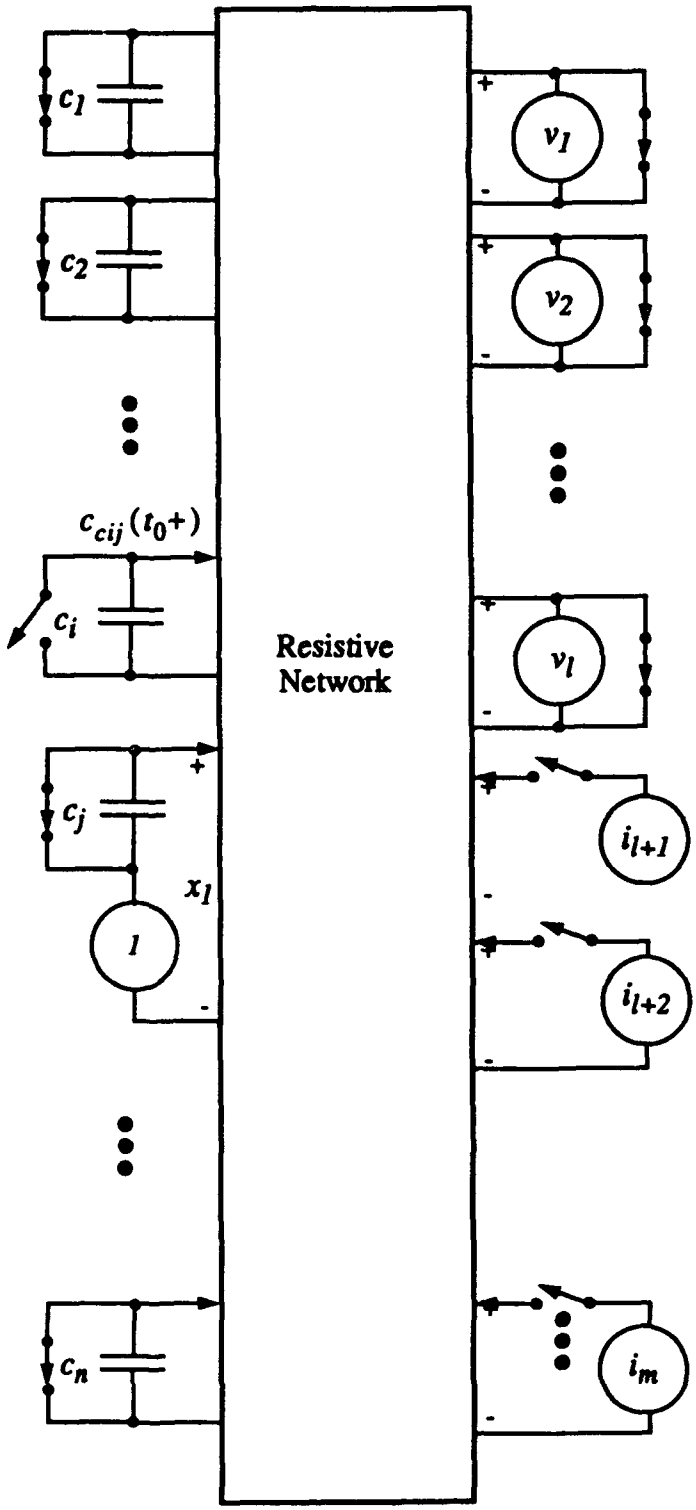


Figure D3 Computation Of The Short Circuit Current In The i'th Capacitor

Under these conditions, Eq. (D5) reduces to

$$\dot{x}_i(t) = \sum_{j=1}^{n} a_{ij} x_j \tag{D8}$$

and further to

$$\dot{x}_i(t) = a_{ij} x_j \tag{D9}$$

giving

$$a_{ij} = -\frac{c_{cij}(t_0+)}{C_i} \tag{D10}$$

Note that computation of the a_{ij}'s require only straight forward and mechanical analysis of the purely resistive network. This leads to ease of conceptualizing, but for large problems other more efficient methods may be desired.

Computation of the B matrix is conducted along similar lines. For each voltage source, V_j, let its value be 1, open it's shorting switch, and leave all other switches in the position of Figure D3. Then

$$b_{ij} = -\frac{c_{cij}(t_0+)}{C_i} \tag{D11}$$

For each current source, i_j, let its value be 1, close its opening switch, and leave all other switches in the position of Figure D3. Then

$$b_{ij} = -\frac{i_{cij}}{C_i} \tag{D12}$$

The C matrix is the observation matrix of temperatures of the conducting material. Each pixel, y, of Y can be considered to be measuring a summation of target surface temperatures, x. Thus the C matrix can be found by elementary analysis of the sensing environment.

The D matrix is the observation matrix of the set of thermal drivers. One can safely set it to 0 for the far IR case. For the near IR case, it is an analyzable function of the thermal drivers.

D4.0 AN EXAMPLE

Assume that a thermally conducting target can be represented by the resistor-capacitor circuit of Figure D4, where the v's are thermal flows from the environment to the target. Then Eq. (D5) becomes

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \\ \dot{x}_6 \end{bmatrix} = \begin{bmatrix} -0.1 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & -0.1 & 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & -0.1 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & -0.1 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 & -0.1 & 0.1 \\ 0 & 0 & 0 & 0 & 0.1 & -0.1 \\ 0 & 0 & 0 & 0 & 0 & 0.1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix} \tag{D13}$$

and the A matrix is Toeplitz and the B matrix is the identity matrix.

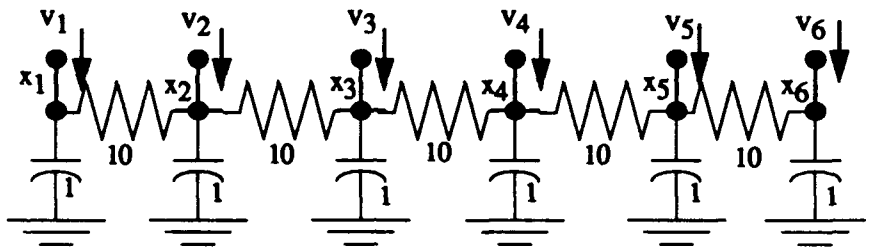


Figure D4 An Example Thermal Network

Now assume that each far IR imagery sensor pixel, y_i, covers x_{2i-1}, x_{2i}, and x_{2i+1} with weights 0.2, 0.6 and 0.2 respectively. Then Eq.(D6) becomes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.6 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0.6 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.2 & 0.6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \tag{D14}$$

The D matrix is the 0 matrix, since in the far IR there is no direct pass-through of the environmental thermal drivers.

Appendix E
THE THERMAL IMPULSE RESPONSE FUNCTION

E1.0 RATIONAL

Appendix C developed the state differential equations for heat conduction networks, implicitly providing a solution for the output, Y.

For cases where the environmental driving thermal forces are not functions of the state variables, X, then it is possible to derive the solution for the output, Y, directly in terms of the input driver vector, V

$$Y(t) = C(t)\Phi(t,t_0)X(t_0) + \int_{t_0}^{t} [C(t)\Phi(t,\tau)B(\tau) + D(t)\varsigma(t-\tau)]V(\tau)\,d\tau \quad \text{(E1)}$$

where

$$\Phi(t) = e^{At} = I + At + \frac{1}{2}A^2t^2 + \frac{1}{3}A^3t^3 + \bullet\ \bullet\ \bullet \quad \text{(E2)}$$

This formulation is key, since in Appendix F it will be found the key to developing the probability of thermal responses to collections of random environmental drivers.

The section below provides further simplification of Eq. (E1) under reasonable ATR target conditions.

E2.0 SIMPLIFICATIONS

$\Phi(t)$ is know in the literature as the fundamental matrix or the transition matrix. For known values of t, it can be calculated to any required precision by computer matrix operations.[1] Other faster numeric methods are based on Sylvester's theorem and the Cayley-Hamilton technique.[2] Laplace transform based symbolic methods include the frequency-domain and the transfer function method.[2] Symbolic mathematics methods include manipulation of the underlying differential equations, and recognition of infinite sums.[3]

In the case of far IR, $D(t)$ is a 0 matrix, giving

$$Y(t) = C(t)\Phi(t,t_0)X(t_0) + \int_{t_0}^{t} C(t)\Phi(t,\tau)B(\tau)V(\tau)\,d\tau \quad \text{(E3)}$$

For heat conduction networks with time-invariant parameters, e.g. metal, this further reduces to

$$Y(t) = C\Phi X(t_0) + \int_{t_0}^{t} C\Phi(t-\tau) B(\tau) V(\tau)\, d\tau \tag{E4}$$

Under normal conditions, there is sufficient thermal interaction between targets and their surrounding environments for $t \gg t_0$ to allow

$$Y(t) \approx \int_{t_0}^{t} C\Phi(t-\tau) B(\tau) V(\tau)\, d\tau \tag{E5}$$

Letting $H(t) = C\Phi(t)$, termed the impulse response matrix of the network, gives

$$Y(t) \approx \int_{t_0}^{t} H(t-\tau) B(\tau) V(\tau)\, d\tau \tag{E6}$$

the matrix convolution solution for the network output, $Y(t)$, in terms of the network input, $V(t)$.

The discrete time version of Eq. (E6) is

$$Y_i \approx \sum_{j=1}^{n} H_{i-j} B_j V_j \tag{E7}$$

Appendix F
DERIVING THE JOINT PROBABILITY DENSITY FUNCTION FOR THERMAL NETWORKS

F1.0 INTRODUCTION

The work of Appendix A through Appendix E culminates in Eq. (E1). Eq. (E1) gives any desired set of thermal conduction outputs as a weighted summation of the amplitude of the thermal input drivers to the target. This appendix extends this deterministic characterizations into the ATR probabilistic realm where the exact history of the targets's thermal environment can not be known.

F2.0 NOTATION

First, the notation must be simplified. For our purposes, let i be the discrete time index corresponding to the time that the IR sensor takes its image. Being well understood, we can drop it from Eq. (E7) to give

$$Y = \sum_{j=1}^{n} H_{-j} B_j V_j \tag{F1}$$

Letting $G_j = H_{-j} B_j$ gives

$$Y = \sum_{j=1}^{n} G_j V_j \tag{F2}$$

Thus Y is seen to be a weighted linear summation of all previous environmental input drivers; that is, all of the individual elements of all V_k vectors.

To make this more obvious and to prepare for further work, let $k = (j-1)m + i$, where m is the size of the V vector. Thus let v_k be the i'th element of V_j of Eq. (F2). This action makes a single input vector out of the many time sequenced input vectors. Upon proper substitution of the g_k, this gives the set of equations

$$y_i = \sum_{k=1}^{mn} g_k v_k \tag{F3}$$

for $1 \le i \le l$, where l is the size of the output vector, Y.

Each output thermal response of a solid is now easily seen to be a weighted summation of all of the instantaneous drivers operating over the body of the target.

Expressing Eq. (F3) in matrix form gives

$$Y = GV \tag{F4}$$

F3.0 A FIRST SOLUTION

Let

$$l = mn. \tag{F5}$$

Then

$$V = G^{-1}Y \tag{F6}$$

, i.e. the inputs are linear functions of the outputs. Let each element of V be written as

$$v_i = g_i(Y) \tag{F7}$$

If the joint probability density function of the set of input thermal drivers, $v_1, v_2, \bullet \bullet \bullet, v_{mn}$, conditioned on a set of target conditions, H, is

$$p(v_1, v_2, \bullet \bullet \bullet, v_{mn} | H) \tag{F8}$$

then the joint probability density function of the set of output target observables, Y, is

$$|J| p(g_1(y_1), g_2(y_2), \bullet \bullet \bullet, g_{mn}(y_{mn}) | H) \tag{F9}$$

where J is the Jacobian of the transformation of V to Y, i. e. the determinant

$$J = \begin{bmatrix} \frac{\partial v_1}{\partial y_1} & \frac{\partial v_1}{\partial y_2} & \bullet & \bullet & \bullet & \frac{\partial v_1}{\partial y_{mn}} \\ \frac{\partial v_2}{\partial y_1} & \frac{\partial v_2}{\partial y_2} & \bullet & \bullet & \bullet & \frac{\partial v_2}{\partial y_{mn}} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \frac{\partial v_{mn}}{\partial y_1} & \frac{\partial v_{mn}}{\partial y_2} & \bullet & \bullet & \bullet & \frac{\partial v_{mn}}{\partial y_{mn}} \end{bmatrix} \tag{F10}$$

F3.1 An Interpretation

Eq.'s (F9) and (F10) form the first truly important results of this research. When the number of time sequenced random environmental driver variables is equal to the number of pixels being observed, then the probability density of the imaged pixels is a simple distortion of the probability density of the joint pdf of the inputs:

- Each axis of the original probability density function is linearly rescaled. As the input variables recede in time from the time of imaging, their axis lengths are reduced.
- Each value of the original probability density function is rescaled by the Jacobian, J, to insure that the output cdf sums to 1.

F3.2 An Example

For illustrative purposes, consider a small scale problem. Let the target region be the thermal conduction network shown in Figure F1.

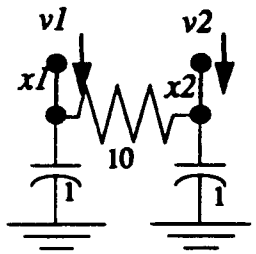


Figure F1 A Small Example Conduction Network

Then the state equation, Eq. (D5)

$$\dot{x}(t) = Ax(t) + Bv(t) \tag{F11}$$

can be given as

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -0.1 & 0.1 \\ 0.1 & -0.1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} \tag{F12}$$

where the v's are environmental driver flows, i.e. currents rather than voltages in the electrical analog of Figure F1.

Taking an IR image at time T, let the two pixels on the target be $y_1(T)$ and $y_2(T)$. Then the state observation equation, Eq. (D6), can be written in discrete time form as

$$y(t) = Cx(t) + Dv(t) \tag{F13}$$

and given as

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{F14}$$

This leads to the impulse response matrix

$$\Phi(t) = \begin{bmatrix} 1-c(t) & c(t) \\ c(t) & 1-c(t) \end{bmatrix} \tag{F15}$$

where

$$c(t) = 0.1t - 0.01t^2 + 0.000666t^3 - 0.0000333t^4 + \bullet \bullet \bullet \tag{F16}$$

$c(t)$ is plotted in Figure F2.

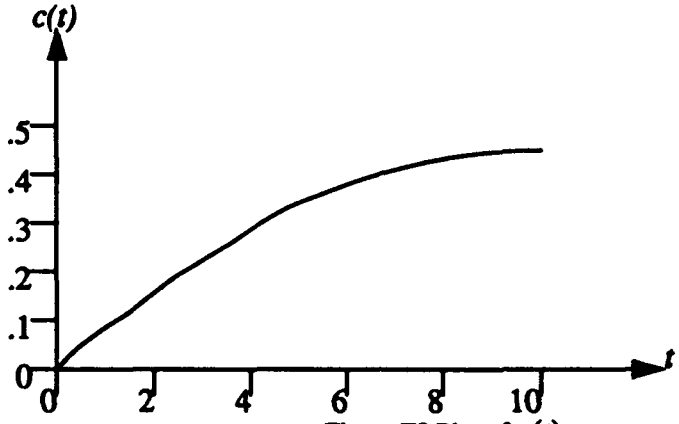


Figure F2 Plot of $c(t)$

Now assume that there are discrete heat inputs at times T_1 and T_2.

$$v(t) = \begin{bmatrix} u_1\delta(t-T_1) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ u_2\delta(t-T_2) \end{bmatrix} \tag{F17}$$

Making appropriate substitutions in the above equations

$$\begin{bmatrix} y_1(T) \\ y_2(T) \end{bmatrix} = \Phi(T-T_1)\begin{bmatrix} u_1 \\ 0 \end{bmatrix} + \Phi(T-T_2)\begin{bmatrix} 0 \\ u_2 \end{bmatrix} \tag{F18}$$

Simplifying

$$y_1(T) = (1-c(T-T_1))u_1 + c(T-T_2)u_2 \tag{F19}$$

$$y_2(T) = c(T-T_1)u_1 + (1-c(T-T_2))u_2 \tag{F20}$$

Solving for u_1 and u_2 gives

$$u_1 = \frac{(1-c(T-T_2))y_1 - c(T-T_2)y_2}{1-c(T-T_1)-c(T-T_2)} \tag{F21}$$

$$u_2 = \frac{-c(T-T_1)y_1 + (1-c(T-T_1))y_2}{1-c(T-T_1)-c(T-T_2)} \tag{F22}$$

From Eq. (F9), the joint probability function of y_1 and y_2 conditioned on H, the event that the IR system is looking at the target of this example is

$$p_{y_1,y_2}(y_1,y_2|H) = |J| \bullet$$

$$p_{u_1,u_2}\left(\frac{(1-c(T-T_2))y_1 - c(T-T_2)y_2}{1-c(T-T_1)-c(T-T_2)}, \frac{-c(T-T_1)y_1 + (1-c(T-T_1))y_2}{1-c(T-T_1)-c(T-T_2)}\middle| H\right) \tag{F23}$$

where $|J|$ is the Jacobian of the transformation

$$|J| = \frac{1}{1-c(T-T_1)-c(T-T_2)} \tag{F24}$$

Visually, Eq. (F23) can be illustrated by Figure F3. The original joint pdf space of u_1 and

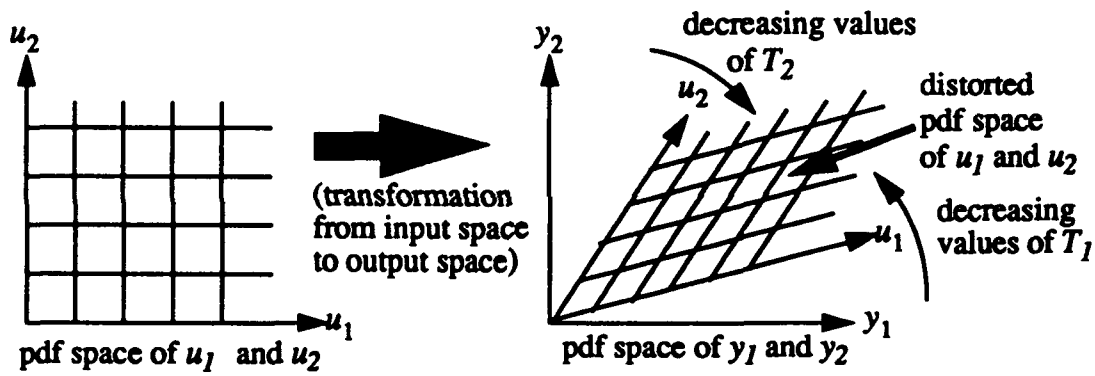


Figure F3 The pdf Space Transformation

u_2, is shown to the left. The joint pdf space of y_1 and y_2 is shown to the right. The thermal conduction transformation from the input of u_1 and u_2 to the output of y_1 and y_2 moves and distorts the pdf space of u_1 and u_2 into the pdf space of y_1 and y_2 as shown. For purposes of illustration, set $T_2 < T_1 < T$.

As T_1 and T_2 approach the value of T, the u_1 axis in the y_1 and y_2 pdf space to the right of Figure F3 approaches the y_1 axis and the u_2 axis approaches the y_2 axis.

As T_1 and T_2 recede into the past away from the value of T, the u_1 and the u_2 axes move in the directions shown by the curved arrows of Figure F3. Ultimately the u_1 and u_2 axes will meet on the $y_1 = y_2$ line, which implies that y_1 and y_2 will then be totally dependent, i.e. $y_1 = y_2$. This is caused by the fact that given sufficient time with no further thermal inputs, the example target of Figure F1 will become constant temperature throughout.

Since the size of the u_1 and u_2 pdf space changes as it is distorted into the pdf space of y_1 and y_2, the Jacobian of Eq. (F23) serves to renormalize the cdf sum of y_1 and y_2 to 1.

The forgoing results are independent of the specific shape of the conditional joint probability density function of u_1 and u_2. However to make the concepts more concrete, arbitrarily choose a uniform pdf for u_1 and u_2 in the shaded region to the left of Figure F4. This is transformed into the uniform pdf for y_1 and y_2. as shown to the right.

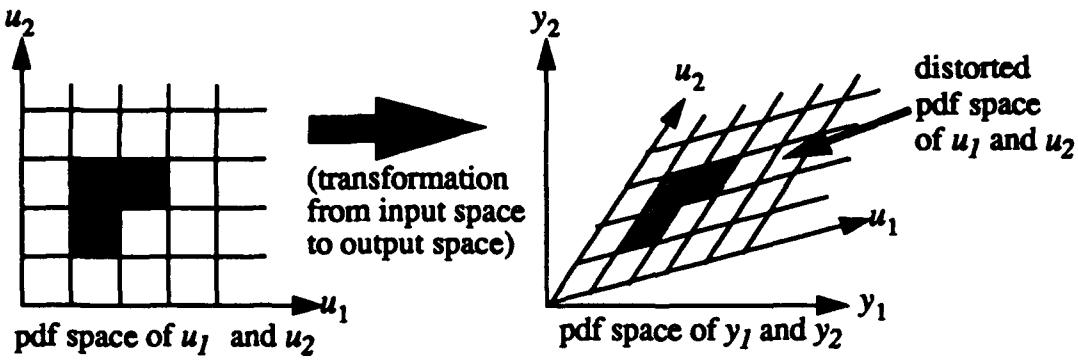


Figure F4 Transform of pdf Spaces

F3.3 Conclusions

One of the major questions which this research into probability modeling of IR target imagery sought to answer is, "Must one know the joint pdf for the environmental drivers in order to determine the joint pdf of the pixels of an IR target?" This appendix shows that for certain cases where the number of pixels equals the number of environmental drivers the answer is "Affirmative." Under these conditions the target is taking a 'motion picture' of its thermal environment. This 'motion picture' continuously decays in amplitude and spatially blurs with time. The pdf of the IR image is thus equivalent to the pdf of this 'motion picture.' This equivalency makes the IR ATR problem both easier and harder.

- The problem becomes harder since one must now probabilistic understand the thermal environment. This will necessitate further research into the time and spatial statistics of thermal drivers in the battlefield (Reference 4 is a first step). It also guarantees that field gathering of statistical parameters of thermal drivers can be used to improve the IR ATR target detection/recognition rate. Example possibilities are statistical collection by the airframe immediately prior to its collection of the IR image of radiant flux levels under the cloud cover and of ground temperatures.
- The problem becomes easier since large regions of the target are normally affected by the same driver values, e.g. solar loading. The state equation formulation of Appendix D lends itself well to incorporation of single drivers acting over regions as well as for points of the target. Also the theory provides a clear theoretical path for considering discrete "large part" modeling of IR targets along the line of Mr. Foley's reasoning at AARA.

F4.0 THE CASE OF MORE THERMAL DRIVERS THAN TARGET PIXELS

The section above considered a special subcase in that the number of thermal drivers, i.e. inputs to the conductive network, was equal to the number of target pixels, i.e. the number of outputs from the conductive network. Since the thermal environment extends arbitrarily far back into the past, the more usual case is the one where the number of thermal drivers exceeds the number of target pixels. Drawing from intuition, one sees also that the significance of thermal drivers recedes as the drivers recede in time. This section is devoted to quantitatively developing the necessary mathematical detail and further insight into the more usual case.

In general, the input joint probability density function conditioned on a set of environmental/target conditions, H, is

$$p(v_1, v_2, \bullet \bullet \bullet, v_{mn} | H) \tag{F25}$$

Now assume l, the size of the output vector, Y., is less than mn, the number of input drivers. Temporarily augment or lengthen the Y vector by setting

$$y_i = v_i \qquad l < i \leq mn \tag{F26}$$

Then by the technique of the section above

$$p(Y) = |J| p(g_1(Y_1), g_2(Y_2), \bullet \bullet \bullet, g_{mn}(Y_{mn}) | H) \tag{F27}$$

or, equally

$$p(Y) = |J| p(g_1(Y_1), \bullet \bullet \bullet, g_l(Y_l), v_{l+1}, \bullet \bullet \bullet, v_{mn} | H) \tag{F28}$$

Now reduce the size of the Y vector by integrating out the temporarily introduced variables

$$p(Y) = \int \bullet \bullet \int |J| p(g_1(Y_1), \bullet \bullet, g_l(Y_l), v_{l+1}, \bullet \bullet v_{mn} | H) \, dv_{l+1} \bullet \bullet dv_{mn} \tag{F29}$$

Of the set of mn drivers, accuracy will be best served by integrating out the time weighted drivers of the smallest amplitude at the time of imaging. These will generally be the oldest.

F4.1 An Interpretation

Eq. (F29) takes the projection of the multi-dimensional pdf from an mn dimensioned space to an l dimensioned space. While this can be computationally complex, the insight is clear. The characteristics of the pdf will stay the same (usually an unlikely situation) or become simpler. This is an answer to the second question of this research, "Is there a law of large

numbers at work which makes complete knowledge of the environmental statistics less important?" The example of the next subsection makes the concept more concrete.

F4.2 An Example

Assume that the example given above in Section F3.2 is modified by adding another driver at time T_3

$$v(t) = \begin{bmatrix} u_1\delta(t-T_1) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ u_2\delta(t-T_2) \end{bmatrix} + \begin{bmatrix} 0 \\ u_3\delta(t-T_3) \end{bmatrix} \tag{F30}$$

where $T_3 < T_2 < T_1$.

Working through the mathematics, this gives

$$y_1(T) = (1-c(T-T_1))u_1 + c(T-T_2)u_2 + c(T-T_3)u_3 \tag{F31}$$

and

$$y_2(T) = c(T-T_1)u_1 + (1-c(T-T_2))u_2 \tag{F32}$$

Now $l = 2 < mn = 3$ and we augment the output Y vector by

$$y_3(T) = (1-c(T-T_3))u_3 \tag{F33}$$

Let the input pdf be given by Figure F5.

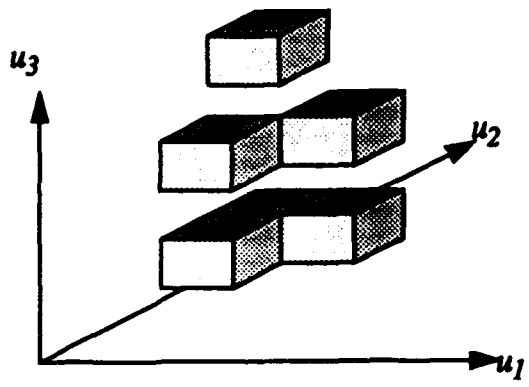


Figure F5 An example Input pdf

This input 3 dimensional pdf is then projected into the 3 dimensional Y space as shown in Figure F6 where the input U space is distorted into the output Y space as shown. The u_1

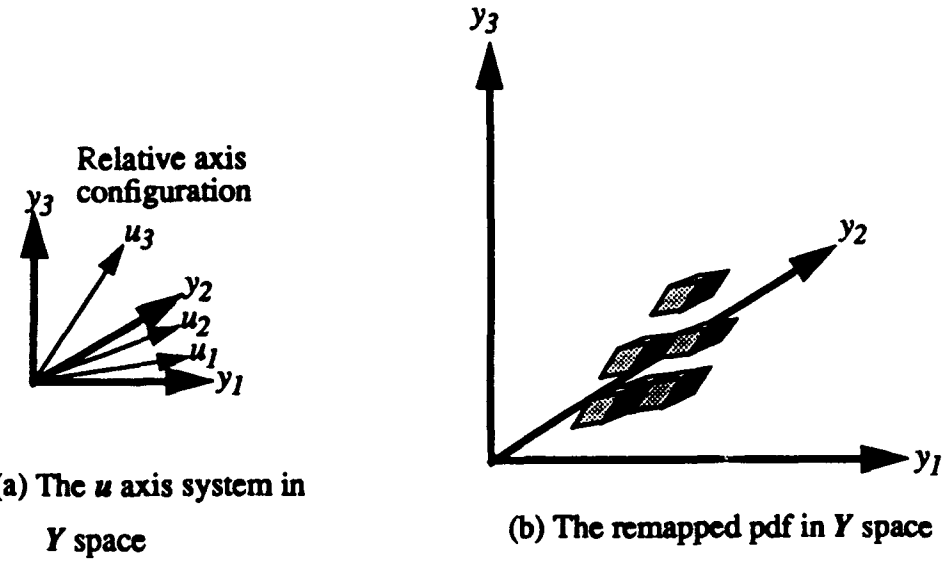


Figure F6 The Remapping Of The Input pdf

and u_2 axes reside in the y_1, y_2 plane. The u_3 axis resides in the y_1, y_3 plane.

When the y_3 axis is integrated out of Figure F6, the distorted pdf function is orthogonally projected onto the y_1, y_2 plane with projection shown in Figure F6.

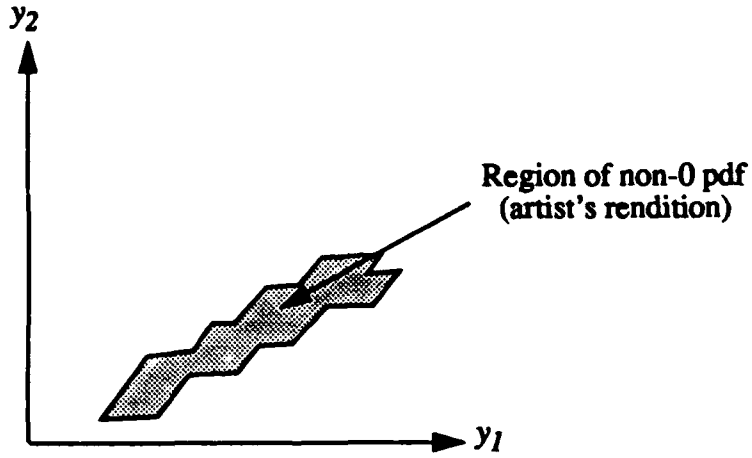


Figure F7 The Projection Of The Input pdf Into The Output Space

The key detail illustrated here is that the complexity of the pdf goes down by the projection process (the pdf of Figure F7 is simpler than that of Figure F5).

F4.3 Conclusions

When there are more environmental drivers than pixels on the IR target region, the joint environmental driver pdf is projected down to a lower space. This reduces the pdf complexity by eliminating any pdf variability along the axes eliminated. If the original pdf function was fractal with dimension d and if e axes were eliminated, then in many cases the projected pdf will have the fractal dimension d-e. It is strongly hypothesized that this reduction of complexity is the enabling mechanism behind whatever success is enjoyed by various ad hoc IR ATR algorithms.

F5.0 THE CASE OF FEWER THERMAL DRIVERS THAN TARGET PIXELS

This case was not considered during this research period since it should seldom occur in practice. If it occurs there are simple extensions to take care of it. One approach can be based on adding sufficient deterministic 0 or nearly 0 valued random drivers to make the number of target pixels and thermal drivers equal. In any case, random variables in the IR sensor would most likely take up any available freedom.

F6.0 OTHER EXTENSIONS OF THE THEORY

- Often there will be poor IR resolution. This can be handled in the state equation approach by setting up an observation vector, Y, that is a weighted sum of appropriate x values.
- The ATR system may need to know the spatial thermal gradient on the target at a particular point on its surface. This can be accommodated in the state equation approach by first setting up two x nodes, x_i and x_2 on the target surface Δd apart along some direction, d. Then form an output y_j value

$$y_j = \frac{x_{i+1} - x_i}{\Delta d} \tag{F34}$$

as an approximation to the spatial partial derivative, $\frac{\partial x}{\partial d}$. By use of the pdf remapping approach, one can then derive the pdf of $\frac{\partial x}{\partial d}$. This approach can be carried further to find total derivatives, etc.

Appendix G
DERIVING THE JOINT PROBABILITY DENSITY FUNCTION FOR THERMAL CONDUCTION NETWORKS WITH NONLINEAR ENVIRONMENTAL COUPLINGS

G1.0 INTRODUCTION

Appendix E considered a thermal conduction network with linear couplings to the environment. Obviously, it may be possible to linearize nonlinear couplings, i.e. radiation and convection, to the environment over the region of interest and continue to use these results. Here the theory is extended to include nonlinear couplings which are so strong as to preclude linearization.

Nonlinear effects add another dimension of difficulty. Indeed, many authors consider that nonlinear differential equations pose an insuperable mathematical problem. The problem of nonlinear probabilistic differential equations is even harder. This appendix sidesteps these difficulties by sufficiently discretizing the underlying problem to the point that it can be solved.

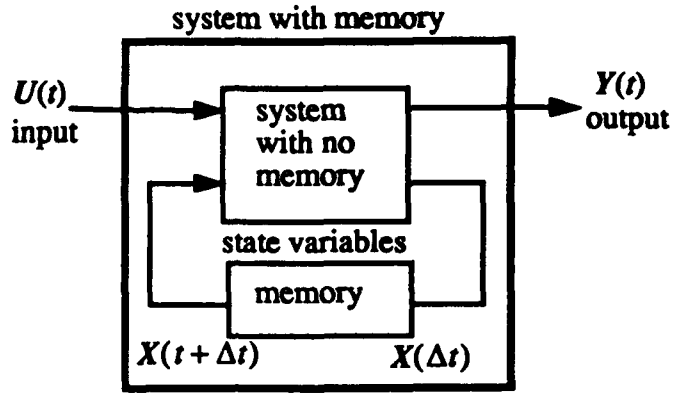
G2.0 THEORY

It is reasonable to assume that the environment is not a function of the target's thermal condition. This makes sense in the ATR case since the environment is much larger that the target. Alternatively, the case where the environment is a function of the target's thermal condition can be solved by extending the target to include the affected portions of the environment.

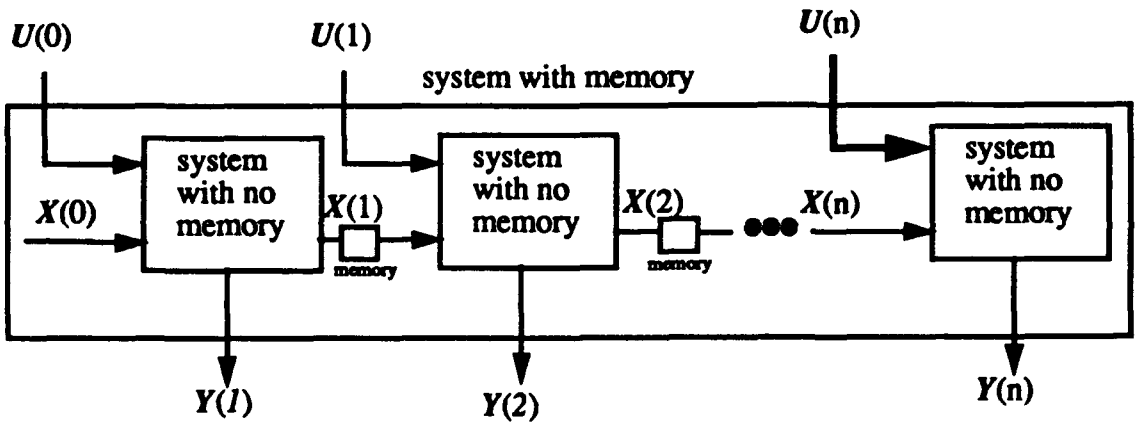
Next, assume that the coupling between the environment and the target is a monotonically increasing function of the difference of specific environmental temperatures and target surface temperatures. (to assume otherwise, would be equivalent to allowing negative resistances in the electrical resistor-capacitor analog - a condition that increases energy transfer as differential environment/target temperatures get smaller)

Nonlinear couplings are equivalent to inserting nonlinear resistances in the state representation of circuits such as Figure D4. These nonlinear resistances can not be handled by the linear state formulation of Eq.'s (D5) and (D6) for the reason that they nonlinearly couple the effects of previous inputs. As a result, the nonlinear couplings void the condition of

linearity on which Appendix D is built, requiring an extended approach to the problem. The approach which is chosen here is discrete time unwarping. Time unwarping is the process of converting a system with memory and sequential inputs into an equivalent system with no memory and paralleled inputs, Figure G1.



(a) Representation of a system with memory



(b) Equivalent representation by a time unwarped cascade of memoryless systems

Figure G1 The Concept Of Time Unwarping

As shown, time is divided into n finite but appropriately small increments. One could potentially pass to the limit as infinitesimal time increments are chosen, although this case is not considered here as it is not needed in the real world of ATR design.

Figure G1b has simplified the mathematics by separating the problems incurred by use of memory and nonlinearity through elimination of the consideration of memory. Now one can solve the problem by working through each nonlinear stage of the cascade from left to right in Figure G1 on a one by one basis. Consider any individual stage, Figure G2. This nonlinear stage contains all nonlinear environment-target couplings present for time *i*. These couplings are function of environmental drivers, U(i), and target surface temperatures, X(i), only. Being thus without memory, the couplings although nonlinear are easily written down in mathematical form once properly modeled. The output, Z(i), then is passed into the linear system box representing the conduction network of the target. The

equations for this linear box can be written in the discrete state equation form discussed earlier in Appendix D.

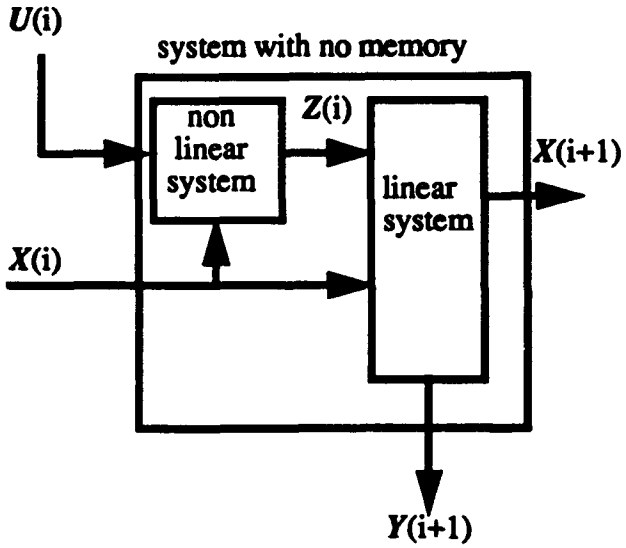


Figure G2 Decomposition Of A Stage Of The Cascade

If the problem was only to get the values of output, the $Y(i+1)$, then the problem would be comparatively easy. The problem; however, is to obtain the pdf of $Y(n)$, where n represents the time of IR imaging. Various subcases are discussed below.

G2.1 The Case of Independent $U(i)$ Over i

If the $U(i)$ inputs are jointly independent over i, the pdf calculation problem is simplified. It is then straightforward to show that for any i, $X(i+1)$ is statistically independent of $U(i)$. This allows the calculation of any stage to proceed as follows. From the pdf of $U(i)$, develop the pdf of $Z(i)$ via very slight extensions of the pdf remapping processes of Appendix F. The extensions required are predominately to take care of the many to one mappings caused by the nonlinearities.

Now form the joint pdf of $X(i)$ and $Z(i)$

$$p(X(i), Z(i)) = p(X(i))p(Z(i)) \tag{G1}$$

The pdf determination of $X(i+1)$ can then proceed exactly as determined in Appendix F, using $X(i)$ and $Z(i)$ as environmental drivers to the linear system.

G2.2 The Case of Dependent $U(i)$

The case where $U(i)$ is dependent over i is a case more likely to physically occur. This is the case where the environment maintains some decreasing correlation over time. The engineering literature seldom treat any case other than Gaussian or to provide more than the first and second moments of more complicated cases. Again wider solution is broadly

considered to be intractably difficult. On the other hand, the full probabilistic solution is needed here for the IR ATR problem.

This more difficult case of non independent environmental drivers may actually be solved as follows. First, one needs the joint pdf

$$p(X(1), U(1), U(2), \cdots, U(n-1)) \tag{G2}$$

Obviously, without a prior thermal characterization of the target, it will be difficult to factor in the effects of $X(1)$ since it depends on its thermal environment up to that time. Appendix B shows, however, that $X(1)$'s effect deteriorates with time so initial assumption errors are reduced as n grows larger. Thus one approach is to choose large n.

Other techniques can also be used to reduce the initial assumption error. Since the absolute time of IR imaging corresponding to n of Eq. (G2) will be known, one can substitute the marginal statistics for $X(1)$ at absolute time corresponding to indexed time 1 based on time of day thermal averages. Prior climate, weather, and cloud cover data can be factored in as well.

In minor extension of the pdf mapping techniques of Appendix F, one can next transform the pdf of Eq. (G2) to that of

$$p(X(1), Z(2), U(2), \cdots, U(n-1)) \tag{G3}$$

by considering the first nonlinear environmental couplings of Figure G2 at time 1.

Then again by the pdf mapping techniques of Appendix F, one can transform the pdf of Eq. (G2) to that of

$$p(X(2), U(2), \cdots, U(n-1)) \tag{G4}$$

thus removing consideration of stage 1 from the cascade of Figure G1. The mathematical process loops from here on until all stages are removed from the cascade. At this point, the output pdf of Y(n) is found from the pdf of X(n).

G2.3 Partial Independence of The U(i)

It is unthinkable that there is no partial independence between the U(i), since total dependence would imply a very high order of dependence of natural events between the various processes of the environment. Markov chains for instance consider only pairwise dependence. The syntax of spoken English is thought to be only of 3- or 4-wise dependence.

If partial independence of the U(i) are observed, say k-wise independence, then the pdf of Eq (G2) can be factored. Each factoring will require only part of the pdf of Eq. (G2) to be subjected to mathematical remapping at any particular stage of Figure G1. This provides corresponding increases in computational efficiency in the mathematical algorithm outlined in Section G2.2 above.

REFERENCES

1 O. I. Elgerd, *Control Systems Theory,* McGraw Hill, 1967

2 P. M. DeRusso, R. J. Roy, C. M. Close, *State Variables for Engineers,* John Wiley, 1965

3 R. J. Schwarz, B. Friedland, *Linear Systems,* McGraw Hill, 1965

4 S. Lashansky, N. Ben-Yosef, A. Weitz, and E. Agassi, "Preprocessing of Ground-based Infrared Sky Images To Obtain The True Statistical Behavior of The Image," *Optical Engineering,* 30, pp1892-1896, December 1991

5 A. Papoulis, *Probability, Random Variables, and Stochastic Processes,* (1965)

6 M. Abramowitz and I Stegun, *Handbook of Mathematical Functions,* Dover Publishing, p376 (1968)

7 R. Cofer, "LADAR Target Detection and Recognition," *Final Report,* USAF-UES Summer Faculty Research Program (1989)

8 R. Cofer, "Probability Determination For Range Imagery Bayesian ATR," *Internal Report,* USAF-UES Summer Faculty Research Program (1989)

9 R. Cofer, "Probability Event Spaces For ATR," *Internal Report,* USAF-UES Summer Faculty Research Program (1989)

10 R. Cofer, "Probabilistic IR Evidence Accumulation," *Final Report,* USAF-UES Summer Faculty Research Program (1990)

11 R. Cofer, "Model Based Bayesian Target Recognition," *Final Report,* USAF-UES Research Initiation Program, (1990)

12 L Kronsjo, *Algorithms: Their Complexity and Efficiency,* John Wiley and Sons, 1979

13 M. Davis, *Computability and Unsolvability,* Dover Publications, 1982

The Enhancement of Connectionist Methods for Recognizing Airplanes from Radar Returns

Final Report on Contract F49620-88-C-0053.

December 30, 1991

Lawrence O. Hall, Principal Investigator
Department of Computer Science and Engineering
University of South Florida
Tampa, Fl. 33620
hall@usf.edu

Abstract

This report covers research into the use of connectionist models to identify airplanes from their radar signature. The radar returns used have been generated by the SRCRCS simulator. In addition a small study has been done on the actual returns from two airplanes. The problem is specified as an unknown aircraft appears on radar and is compared by a connectionist model to known aircraft radar signatures. The study has primarily examined recognizing planes as their aspect angle changes. A 90 degree aspect angle spectrum has been studied. The ideal situation is to have the aspect angles in the train set far apart for a smaller training set. The major problem becomes one of representing the data to the connectionist network. The proper representation, or feature extraction from the data, allows good recognition to occur. Several methods of encoding the radar returns for recognition are studied. Two types of connectionist networks are used in this study, a feedforward backpropagation network and an instance-based learning hybrid connectionist network (SC-net). It is shown that the use of self configuring forms of backpropagation networks are needed in this type of problem. Recognition rates of better than 90% with training at every 5° of aspect angle have been achieved. Aspect angle recovery to within an average of 3° of the actual aspect angle has been achieved on correctly recognized planes.

1 Introduction

The research described here involves recognizing airplanes from their radar signatures. The goal is to use air to air radar because of its long range potential and all-weather capability. Test results have shown that the returns from a high range resolution (HRR) radar can uniquely describe targets. Returns from HRR radar represent illuminated targets as range profiles along the line of sight of the radar. These range profiles often vary appreciably over small changes in aspect angle. The changes can be attributed to the illumination of different scatterers and/or a compression of inter-scatter distances due to a change in the geometry of the problem, though the latter is a slight effect. Since the HRR data is so aspect-dependent, target recognition with all necessary aspects currently requires a very large database. The recognition task is further complicated by the fact that the radar returns must be aligned with the stored data for comparison. The issues discussed above motivate the need to develop an efficient/compact method of representing and comparing HRR data.

Supervised connectionist techniques are used as the means to recognize the planes from radar signatures. Their generalization capabilities are exploited to limit the training set to a *manageable* size. Feedforward backpropagation networks in the form of Quickprop [5] are experimented with unsuccessfully. The instance-based connectionist learning system embodied by the hybrid symbolic, connectionist SC-net system is used primarily. Several other learning systems are explored briefly in this research. They consist of two connectionist models that grow their own structure, Cascade Correlation [6] and Divide and Conquer Networks [20].

There were seven airplanes studied in this research. They consisted of models of the F4, F14, F15, F16, F18, T38 and Lear, as generated by the SRCRCS tool developed by the Syracuse Research Corp. The data simulates a wide band width and high range resolution radar (HRR) with a 20 meter range window, consisting of 256 discrete points. The magnitude of the return at each point corresponds to the radar cross section at a specific range on the

aircraft. The planes were studied at aspect angles from 0 to 90 degrees, where 0 degrees indicates a plane coming directly at the viewer. This range was felt to be sufficient to determine whether our approach would be viable for actual planes. It covers the space of views that might be seen of an approaching aircraft. Since, the models were symmetrical, only up to plus 90 degrees was studied.

A small amount of real data was available to us on this study. One type of aircraft included examples with 2 different types of armaments. There were 812 measured points in the return. We had 21 returns of high range resolution radar from 3 aircraft of 2 different types. The aspect angles were in a 3° range around 180° meaning the planes were going directly away from the radar site. One issue that this data forced us to acknowledge was the following. The planes were not necessarily centered in the window of the radar return as they were with the generated returns. This fact about real data shaped all of the methods we devised to represent radar returns to our learning systems. The problem of representation for learning is clearly much simpler in the case that it is known where the center of the airplane return resides. There is less chance of encoding noise as a feature for identification, which would lead to widely varying and inaccurate results.

The most difficult problem in our view is determining what features to use for the connectionist models to learn to make their identifications. A radar return consists of a set of peaks which result from scatterers on the airplane. There will be noise in the return, which can provide false peaks. The first task is to exclude as much noise as possible without losing real information. This involves establishing a noise floor as a basis from which to measure the height of peaks. The problem then, is how to turn the peaks into meaningful features from which accurate recognition can be done. It is our contention that if this is accurately done, many learning algorithms may be able to generalize effectively from a well-chosen training set.

The method of recognizing airplanes from radar returns, described here, is not solely

based upon training a connectionist network and using the trained network for recognition. It also depends upon encoding the radar returns for training and testing. From one return a number of patterns are created which are then used for training. Given a return for testing, the return is used to generate multiple patterns in the same manner as for the training phase. The patterns are then presented to the trained network and the plane is identified as the one which has its output turned on for the most patterns.

In the following, a description of the main learning algorithms employed is given, then the crucial topic of how to represent the returns as features to the algorithms is discussed, the actual recognition algorithm is discussed, then the results from experiments with different representations schemes is presented and finally a summary of our research findings.

2 The connectionist learning algorithms

The algorithms described below are supervised learning algorithms which require the examples in the training set to be labeled or classified. Quickprop requires a fixed architecture be chosen by the user of the system, while SC-net configures its connectionist architecture based upon the training examples presented it. Cascade Correlation is a connectionist algorithm that uses the Quickprop training method, but grows its own structure based on the training set presented it.

2.1 Quickprop and Cascade Correlation

The Quickprop algorithm is a generalization of the Backpropagation algorithm, which is significantly faster during training than Backpropagation [5]. It can be up to an order of magnitude faster. Further, comparable results have been observed [14] when compared against the backpropagation trained network.

Cascade correlation begins with just input and output layers. Every input is connected to every output, and all the connections have adjustable weights. Moreover, there's a bias

input, permanently set to +1. The Cascade architecture is shown in Figure 1. The output nodes may just produce a linear sum of their weighted inputs, or they can use some non-linear activation function. As the network is trained, new hidden nodes are added to the network, one by one as needed, and as dictated by the learning algorithm. Cascade-correlation starts with no hidden nodes, and using Quickprop, it trains the direct connections between the input and output nodes over the entire training set until no significant error reduction is achieved. Then the error is computed by running the network over the whole training set. If the error is less than ϵ (set by the user), the algorithm terminates; otherwise, a new hidden node is added in an attempt to eliminate or reduce the residual error. The inputs of the new node come from every input node and all other hidden nodes in the network; each hidden node adds a new one-node "layer" to the network. Initially, the outputs of the new node are not yet connected to the active network. A number of passes are run over the training patterns, adjusting the weights of the input connections to the new nodes after each pass. This adjustment is done in the direction that maximizes S, the sum over all output nodes of the magnitude of the correlation between the candidate node's output value z as shown in Figure 1; and the residual output error, E_o, observed at each output node o.

Finally, quickprop is used for a fast-converging ascent to maximize S. When the magnitude of the correlation stops improving, the new candidate node is installed in the active network; its outputs are now connected to the output nodes. Then its input weights are frozen, and its output weights are trained to minimize the network's output error. This is done in the same way the direct connections were trained, using quickprop. If the network's performance is not satisfactory and no significant error reduction is achieved, a new hidden node is introduced; the cycle is restarted. This cycle is repeated until the error $E_{cc}(\mathbf{w})$ becomes acceptably small (or until we give up). In addition, cascade-correlation can use a pool of candidate nodes instead of just one node. These nodes are initialized with different random weights; the candidate with the best correlation is installed. Some implementation improvements can also

be found in [6].

2.2 Divide and Conquer Networks

The algorithm described here differs from other self-configuring algorithms that make use of a backpropagation-like training method in that it will add cells in multiple hidden layers unlike the algorithms reported by Ash and Hirose, et.al. [1, 11]. However, it can add multiple nodes on each hidden layer unlike Cascade Correlation [6] and does not make use of any correlation measure. Further, divide and conquer networks use only the examples that have not already been correctly classified in the incremental training process. The algorithm was partly inspired by the difficulty in learning to recognize planes with Backpropagation or Quickprop. The algorithm is new and relatively untried. It is completely described in Appendix A.

2.3 SC-net

SC-net is a hybrid symbolic, connectionist learning system [9]. Rules may be directly encoded into the network [18]. A limited form of variable binding is supported. After learning is complete rules may be generated from the network. The learning algorithm is instance-based and causes the structure of the connectionist network to be configured based upon the examples presented to SC-net in the training phase. Only the inputs and the number of outputs are fixed initially. The learning algorithm is also incremental which allows examples to be presented one at a time or in batches without requiring retraining on examples seen earlier.

The learning algorithm works as follows. For each labeled example presented to the system a forward pass is made. There are three possible outcomes for each output.

1. The example is recognized correctly.

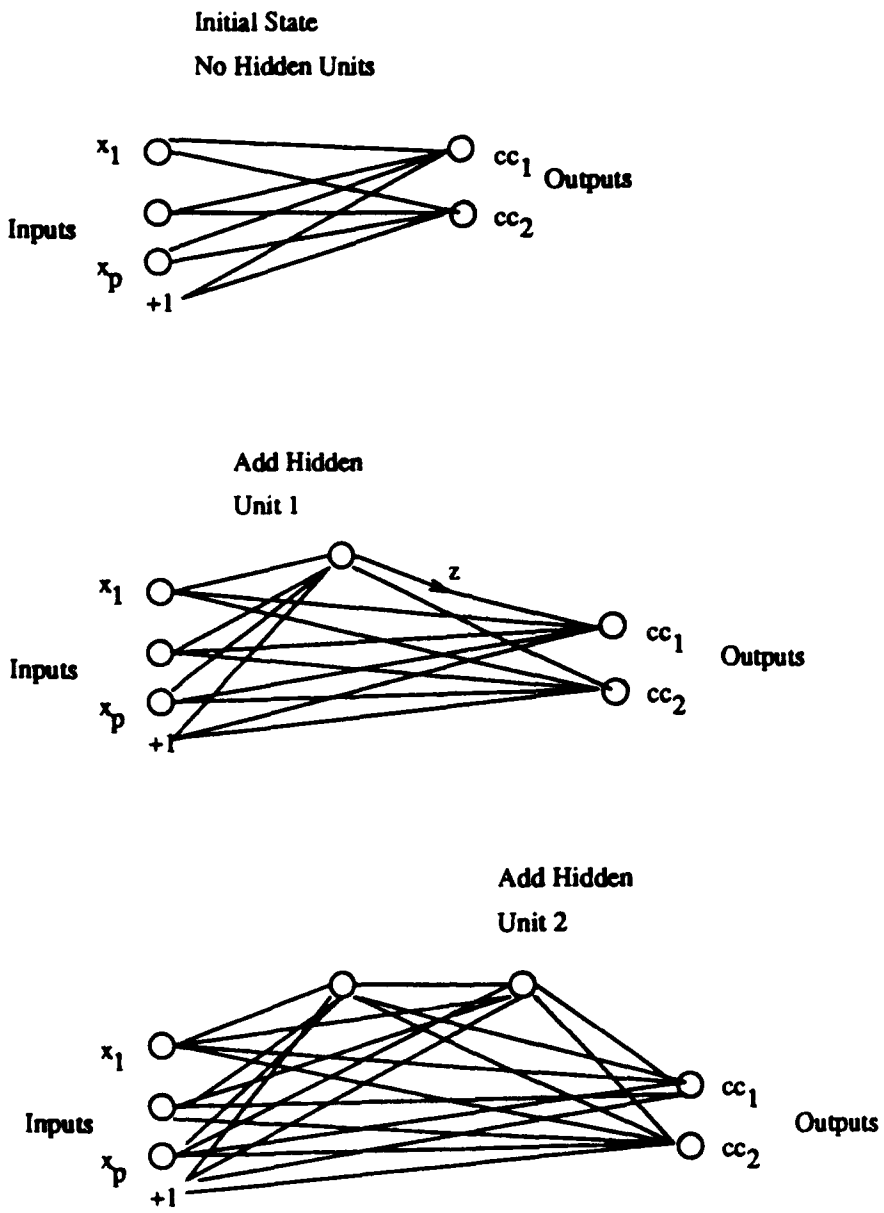


Figure 1: The Cascade architecture from initial state through the addition of two nodes. The weights to the hidden nodes are frozen when the hidden nodes are added.

2. The example is not recognized correctly, but is within a threshold of the desired output value.

3. The example is not correctly recognized.

In the first case, no action is taken. In the second case the bias of a cell in the existing network will be modified to enable the system to recognize the new example while retaining the ability to recognize previously seen examples. In the third case, the network structure will be added to, in order that the training example may be correctly recognized the next time that it is presented to the network. This is called the recruitment of cells phase.

As an instance-based algorithm, only the examples seen in the train set are recognized unless a form of post generalization is used. With a nearest neighbor classifier, for example, a distance metric is commonly used to classify examples in the test set which do not exactly match the training examples. SC-net has two forms of post training generalization [18]. The simplest to describe, and the one used in this work, is the min-drop feature. When a test pattern is presented to the system, which has not been seen in the training set no output is turned on. Hence, the min-drop is applied to find the nearest corresponding output for the example, if one exists. The min-drop feature works by dropping a pre-set number of inhibitory connections in the network one at a time until an output is turned on or the limit of the connections allowed to be dropped is reached. It provides surprisingly good generalization. The other method is a formalized covering algorithm which removes permanently connections that are not needed for recognition of the training examples.

SC-net allows the use of fuzzy variables [18] to represent inputs in continuous data. This feature was used in the work reported here and is important in the following sense. The return values in db are real valued and must be encoded in some fashion to SC-net which accepts inputs between 0 and 1 only. The conversion to [0,1] is done in the binning process when normalization occurs. However, this work depends upon the different values represented by the bin magnitudes being meaningful (if compressed). Further, it has been found in other

Table 1: Fuzzy partitions used for bin values

p1	0 - 0.1
p2	0.1 - 0.2
p3	0.2 - 0.3
p4	0.3 - 0.4
p5	0.5 - 0.6
p6	0.6 - 0.7
p7	0.7 - 0.8
p8	0.8 - 0.9
p9	0.9 - 1

research that encoding real valued features into a set of features each of which is turned on when the original value is within its range can be an effective aid to the learning process [3, 4].

In these experiments, it was unknown how to partition the interval into the proper sub-intervals so we used those shown in Table 1 and dynamic plateau modification (DPM) to modify the fuzzy regions [18]. The use of fuzzy variables means that for each input feature, 9 fuzzy values will be generated for use in training. In Figure 2, the membership function for p2 is shown. The membership value for inputs is 1 in p2 when the input values lie in [0.1,0.2]. Membership in p2, $\mu_{p2}(x)$, has a linearly declining value as the actual value goes from 0.1 towards 0 or from 0.2 towards 1, respectively. DPM learns to modify the arms of the membership function for each of the respective intervals by bringing them in or forcing them to 0 earlier. For example, DPM might cause $\mu_{p2}(x)$ to go to 0 at 0.3 instead of 1, shown by the dashed line in Figure 2.

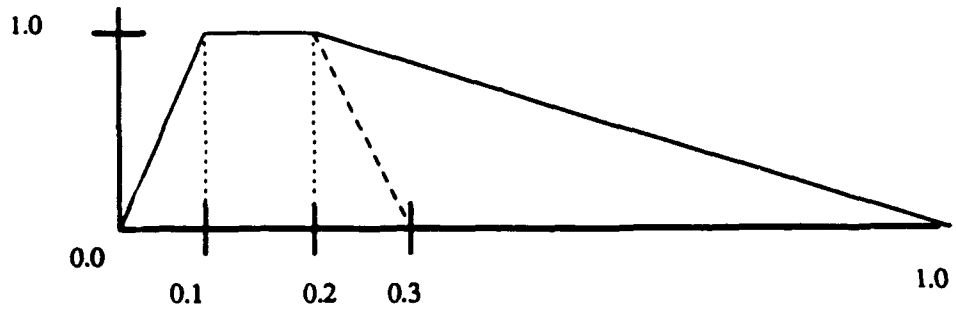


Figure 2: The initial fuzzy membership function for p2.

3 Representations of the Radar Returns

The radar returns in this study are a set of points representing a radar cross section at separations of about 8cm with each point having a value in the decibel (db) range. The choice of features for learning from these points is problematic for the following reasons. Each point of the return will not correspond to the same point on the same plane at the same view, in general. The magnitude of the returns is not guaranteed to be the same for different returns on the same plane at the same view. Noise can have unpredictable effects on the returns.

Taking the magnitude of the return at each point will not provide a viable train set. The second possibility is to use the relative magnitude of each point's return value, but unless the plane can be reliably positioned in the return window this approach will not provide a viable train set either. Further, it is desirable to have the smallest number of viable features for training. Hence, as others have, we incorporate the concept of grouping a number of points in a return into bins [7]. The size of the bins, or number of return points each contains, can be chosen at training time.

After choosing a bin size for training, the biggest problem is to ensure that the bin from the same portion of the aircraft is provided to the same input of the learning algorithm for each training pattern. That is, if a bin incorporates for example the first 8 points of a return from a plane and is mapped for one pattern to the first input or feature of the connectionist

network, all future patterns involving that plane at the same aspect angle would ideally have the same bin mapped to the same input. However, if you could identify the center or start of the plane in the return without ambiguity you would probably not need to use a connectionist system for recognition.

Three algorithms for representing the radar returns to the connectionist networks have been developed in an attempt to mitigate the difficulties discussed. Each puts groups points in the return into bins of a user set size. The bins are represented by the average value of heights of the peaks within them. Each algorithm creates more than one training pattern from each radar return by creating different alignments of the bins. The idea is that the group of patterns for a return (each labeled with the same plane) will be recognizable. Hence, during testing the same type of group will be created and the class assigned will be the one which is turned on by the largest number of patterns from the group.

The first algorithm is called A.1 and is shown in Figure 3. It aligns a return by the k largest bins, where the magnitude of a bin is calculated as the average height of the peak values contained within. Hence, k patterns are created for each return such that each of the of the k largest bins is centered to make the individual patterns. To prevent "important" bins from being shifted out of the set of features, left and right "guard bins" are created and they contain the largest bins that would be shifted out (when centering the the set of features to the left and right respectively. The guard bins serve as the left or rightmost bin depending upon the direction of shift. When shifting in one direction, bins with an average value of 0 are inserted as the outer bins in the opposite direction. Hence, shifting must be limited as the bins shifted in contain no information. The number of inputs or features provided to the learning algorithm for A.1 will be the number of bins.

The second algorithm is called A.2 and is shown in Figure 4. It is also an attempt to use shifting of bins to capture a return in several patterns. This time 2j+1 training (testing) patterns are created from a return. The parameter j is chosen by the algorithm user. A.2

The Binning Algorithms

Binning Algorithm A.1
(Align by k largest Scatterers)

(0) Inputs: Bin-Size, k (k < #Bins)
(1) Determine largest and smallest scatterer
(2) Calculate average scatterer height
(3) Remove noise (Cutoff is average scatterer height)
(4) Create Bin-Vector of size Bin-Size
(5) For i=1 to k do
  (5.1) Align by ith largest bin (center bin)
  (5.2) Create Left-, and Right-Guard-Bin
  (5.3) Set Left-Guard-Bin to largest bin shifted out to the left (if appropriate)
  (5.4) Set Right-Gurad-Bin to largest bin shifted out to the right (if appropriate)
  (5.5) Print Bin-Vector
(6) Halt

works by creating bins and centering each of the j bins to the left of the center of the actual return and each of the j bins to the right of the actual return. The center bin of the return is also the centerpoint for a train/test pattern. Again guard bins are used on the left and right sides to capture the largest bins that would otherwise be shifted out of the set of features to be presented to the connectionist network. For n bins there will be n features, one for each bin.

The third algorithm is called B and is shown in Figure 5. In this algorithm no information is shifted out of the training/testing pattern. This is accomplished by creating a temporary "bin" vector of twice the number of features as the number of bins generated from the radar return. Now as in A.1 the k largest bins are centered, but because they are centered in a vector of twice the number of actual bins all information is retained. The temporary bin vector for each train/test pattern is then compressed by allowing neighboring bins to become one new bin with a magnitude that is the average of the average peak heights in both bins. This results in patterns with k, the number of original bins in the binned return, features.

Binning Algorithm A.2
(Align by j Left-Right Shifts)

(0) Inputs: Bin-Size, j (j < int(#Bins/2)+1)
(1) Determine largest and smallest scatterer
(2) Calculate average scatterer height
(3) Remove noise (Cutoff is average scatterer height)
(4) Create Bin-Vector of size Bin-Size
(5) For i=center-j to center+j do
  (5.1) Align by ith element of Bin-Vector (center by this element)
  (5.2) Create Left-, and Right-Guard-Bin
  (5.3) Set Left-Guard-Bin to largest bin shifted out to the left (if appropriate)
  (5.4) Set Right-Gurad-Bin to largest bin shifted out to the right (if appropriate)
  (5.5) Print Bin-Vector
(6) Halt

Figure 4: Algorithm A.2

Binning Algorithm B
(Align by k Left-Right Shifts, no shifting out)

(0) Inputs: Bin-Size, k (k < #Bins)
(1) Determine largest and smallest scatterer
(2) Calculate average scatterer height
(3) Remove noise (Cutoff is average scatterer height)
(4) Create Bin-Vector of size Bin-Size
(5) Create Bin-Vector-2 of size 2*Bin-Size
(5) For i=1 to k do
  (5.1) Align by ith largest bin (center bin within Bin-Vector-2, no shifting out)
  (5.2) Compress Bin-Vector-2 into Bin-Vector by taking neighboring bins and compressing them into one and taking the average value for that new bin
  (5.3) Print Bin-Vector
(6) Halt

Figure 5: Algorithm B

4 Recognizing airplanes from radar returns

Every radar return for both training and testing will be pre-processed by the chosen binning algorithm. During the testing process all the patterns generated by the algorithm will be presented to the trained connectionist network. The plane recognized is the one that is recognized the most times given the patterns. That is, if there are 4 patterns and the output of plane A is turned on for 2 of them, the output of plane B for 1 and the output of plane C for 1; plane A will be presented as the plane recognized for the given radar return. In the case that a tie exists, no decision will be made.

The major factor in getting a good recognition percentage would seem to be the choice of data encoding algorithm and then of the number of bins and the amount of shifting that is done. In the experimental results section, we explore the effect of the choice of parameters on the accuracy of recognition.

For Quickprop the other issue is how to choose the number of units in the hidden layers and number of hidden layers. In general with many different configurations, convergence was not accomplished. This indicates that it is a hard problem in the sense of choosing a network configuration that will work reasonably well. Training for up to 2 weeks on a SUN Sparcstation was attempted. This difficulty lead us to develop the divide and conquer neural network which grows its own internal configuration based on the training data presented it.

5 Experimental Results

The results reported here are with the use of SC-net for training and testing, unless otherwise noted. Despite our best efforts, we could only get Quickprop to converge for the airplanes from 0-45 degrees aspect angle by creating 3 patterns for each return by centering the largest magnitude bin and shifting left and right once. This means that we had a smaller number of training patterns than in the experiments described below. It is believed that the same

sort of problem would impede any algorithm for this domain that requires a pre-specified architecture for learning. Hence, we do not show results from any algorithm that is not self-configuring.

In the following, we will discuss results with the SRCRCS generated returns first. The experiments with recognizing the planes from 0 to 90 degrees of aspect angle shift are first described. Next a discussion of the limited results from the small amount of actual data is provided. Finally, a discussion about recovering the aspect angle of the recognized plane comes next.

5.1 Recognition in the 0 to 90 degree aspect angle range

There are 7 planes each with simulated radar returns at aspect angles from 0 to 90°. Hence, a total of 637 returns exist. Of these SC-net is trained on the patterns generated from those 5° apart beginning at 0, or 0, 5, 10, 15,...,85, and 90 degrees. Hence, 133 returns are used for training, which leaves 504 returns unseen by the learning algorithm.

Table 2 shows results using encoding algorithm A.1. The values shown are the percentage correct for the bin size and number of patterns or large bins centered. The bin size is in terms of the number of points in a return that are in a bin. Where the division is uneven the last points of a return will not be placed in a bin. TC stands for the total percentage correct ($1 - \frac{Total\ misses}{637}$) and UC stands for the percentage correct on the unseen returns ($1 - \frac{Total\ misses}{504}$) to provide an explicit idea of how able SC-net is to generalize. It is clear that performance increases with the number of patterns generated from a return. For example in the case that 6 patterns are generated with the use of the 6 bins with the largest magnitudes, the average performance on the radar returns not in the training set is 89% versus 76% with 2 patterns and 69% with 1 pattern.

Table 3 shows the results with encoding algorithm A.2. The results are presented in the same format as before, but now only with 1, 3 and 5 patterns generated. It can be seen that

Table 2: Results with algorithm A.1.

Bin Size	1 Pattern		2 Patterns		3 Patterns		4 Patterns		5 Patterns		6 Patterns		9 Patterns	
	TC	UC	TC	UC	TC	UC	TC	UC	TC	UC	TC	UC	TC	UC
3							0.82	0.77	0.88	0.85	0.91	0.89		
4									0.90	0.88	0.91	0.89	0.93	0.91
5	0.74	0.67	0.78	0.72	0.85	0.81	0.89	0.86	0.91	0.89	0.94	0.92	0.94	0.93
6	0.75	0.69	0.81	0.76	0.86	0.82	0.88	0.85	0.92	0.89	0.92	0.90	0.94	0.93
7	0.78	0.72	0.79	0.73	0.88	0.85	0.89	0.86	0.91	0.89	0.92	0.90	0.94	0.93
8	0.79	0.73	0.73	0.66	0.87	0.84	0.89	0.86	0.90	0.88	0.91	0.88		
9	0.76	0.69	0.75	0.68	0.83	0.78	0.89	0.86	0.87	0.84	0.88	0.85		
10	0.72	0.65	0.74	0.67	0.86	0.82			0.87	0.83	0.87	0.83		
11	0.78	0.72	0.73	0.66	0.78	0.72								
12	0.77	0.71	0.74	0.67	0.74	0.67								
13	0.78	0.72	0.76	0.70	0.79	0.73								
14	0.75	0.68	0.71	0.64	0.80	0.75								
15	0.74	0.67	0.71	0.63	0.79	0.73								

the results with 1 pattern are identical to those for A.1 with the exception of a bin size of 7 for which the recognition difference is minor. This points out some of the regularity found in the SRCRCS data. With 3 and 5 patterns A.2 is outperformed by A.1 by several percent. It appears that focusing on the largest scatterers, which is what A.1 does is a better strategy than simply doing left-right shifts of the data.

Table 4 shows the results using encoding algorithm B. It provides results comparable to A.1, as it is not losing any data. Again the best performance comes with the use of more patterns with 8, 9 and 10 being the best number of patterns for this algorithm. It is the case that this algorithm works better with a smaller bin size. A smaller bin size will mean that more processing is necessary, because there are more inputs to the network.

A modification to algorithm B in which the number of bins centered to create training and testing patterns (k in algorithm B) is based on a threshold has also been tried. In this scheme all bins with a magnitude (which is the average value of the returns captured by it) above the threshold T are used as centerpoints to the different patterns generated from one return. Now

Table 3: Results with algorithm A.2.

Bin Size	1 Pattern		3 Patterns		5 Patterns	
	TC	UC	TC	UC	TC	UC
3						
4						
5	0.74	0.67	0.82	0.77	0.82	0.77
6	0.75	0.69	0.81	0.77	0.86	0.82
7	0.75	0.69	0.83	0.79	0.87	0.84
8	0.79	0.73	0.84	0.79	0.85	0.81
9	0.76	0.69	0.80	0.74	0.83	0.79
10	0.72	0.65	0.82	0.77	0.84	0.80
11	0.78	0.72	0.80	0.74	0.84	0.79
12	0.77	0.71	0.73	0.66	0.81	0.76
13	0.78	0.72	0.82	0.77	0.86	0.82
14	0.75	0.68	0.80	0.75	0.82	0.77
15	0.74	0.67	0.76	0.70	0.82	0.77

Table 4: Results with algorithm B.

Bin Size	3 Patterns		5 Patterns		8 Patterns		9 Patterns	
	TC	UC	TC	UC	TC	UC	TC	UC
3	0.82	0.77	0.92	0.90	0.93	0.92	0.94	0.93
4	0.83	0.78	0.90	0.87	0.93	0.91	0.94	0.93
5	0.76	0.70	0.89	0.86	0.90	0.87	0.91	0.89
6	0.77	0.71	0.83	0.79	0.87	0.84	0.86	0.83
7	0.75	0.68	0.83	0.78				
8	0.77	0.71	0.83	0.78				
9	0.79	0.74	0.81	0.76				

each return may generate a different number of patterns. The higher T is the less patterns that will be generated. Experiments were tried with $.10 \leq T \leq .30$. varying T by 0.01. Table 5 shows a set of representative experimental points with a bin size of 5. The number of patterns linearly increases as T decreases. The number of misses decreases as T decreases although $T_1 < T_2$ does not guarantee less misses with the use of the patterns generated by T_1. The trend is downward, though. The table clearly indicates that generating more patterns based on the highest peaks in the return provides better recognition performance.

5.1.1 Learning time and other algorithms

The learning time was an average of about 5 hours on a SUN-3 workstation with SC-net for the cases where multiple patterns from one return were used. There were 1197 training patterns for the case of 9 patterns per return which appeared to be the best case for learning using encoding algorithm A.1 or B. This caused SC-net to recruit 1197 cells for the recognition network.

Table 5: Results with algorithm B and varying threshold.

T	Pattern count	Misses
.30	2621	75
.25	3133	73
.20	3775	53
.16	4461	49
.10	6009	38

To provide a perspective to this work, we used Cascade Correlation on algorithm B with a Bin size of 4. It also grows its own structure, but is somewhat more of a conventional connectionist system. It recruited 86 hidden units, ran for 24 hours on a Sparcstation 1+, had residual error of 2.6168 and an accuracy of 77% on the entire set of patterns (with a threshold of 0.6). Again a the plane with the largest vote after all patterns generated by algorithm B for a return are presented is the plane classified. It mostly confused the lear jet model with other planes, primarily the T38. Several higher thresholds were tried to force the algorithm to have less residual error in the training, but performance did not improve.

The divide and conquer network (DCN) approach has been tried on the problem in which the radar data is not encoded at all. Training is done at 5° intervals and testing is done on the 7 planes at all of the other aspect angles. DCN provided an accuracy of 71.2% vs. 74.2% for SC-net and 65.9% for Cascade Correlation.

5.2 Results from actual data

The ARTII data consists of 812 return points. The 2 planes in the study are likely significantly shorter than the range. Hence, alignment of the return to the connectionist network is crucial. The elevation angle of the planes is near 0. The aspect angles are within a 3° arc of 180°

Table 6: Results from actual radar data.

Bin Size	Algorithm A.2.1	Algorithm A.2.2
25	.95	.81
20	.90	.86
15	.86	.86

essentially a view of the planes receding or looking at the tail of the planes. There are significantly more returns in a smaller aspect angle range with this data. In a five degree range, we would use a maximum of two returns per plane for training. Here, we have been even more restrictive and used only 2 returns (one of each type of plane) for training with the remaining 19 used for testing. With such a small data set no definitive conclusions can be drawn, but the results of generally getting 80% or greater recognition are encouraging.

Variations of algorithm A.2 were used to encode the returns into patterns. Let us call the variations A.2.1 and A.2.2, respectively. They both involve centering the highest bin or making that bin the center bin in the feature vector to be presented for learning. A.2.1 then implements algorithm A.2 with j=1 or 1 left, right shift to create 3 train/test patterns. A.2.2 implements A.2 with j=2 or 2 left, right shifts to create 5 train/test patterns. It is necessary to center the returns by the highest bin to locate the plane within the original radar return. The plane may be in the leftmost, rightmost or center portion of the actual return with noise elsewhere and we, of course, have no control over the original location.

Table 6 shows the results of 3 experiments using the same two returns for training. The accuracy is calculated over all 21 planes. The number of misses can be calculated by taking the nearest integer value of the number of planes times one minus the accuracy value. Three different bin sizes are used. Algorithm A.2.1 does better with larger bins, while A.1.1 does better with smaller bins.

```
Assume the following K=9 diagnoses have been made:

(1) Plane F4 at 5 degrees
(2) Plane F4 at 10 degrees
(3) Plane F4 at 5 degrees
(4) Plane F4 at 5 degrees
(5) Plane F16 at 20 degrees
(6) Plane F18 at 5 degrees
(7) Plane F4 at 15 degrees
(8) Plane F4 at 0 degrees
(9) Plane F18 at 10 degrees

Plane F4 wins the vote (6/9) and the predicted or recovered aspect
angle is:
  (5+10+5+5+15+0)/6=6.67
```

Figure 6: An example of calculating an aspect angle from a test return.

5.3 The recovery of aspect angle

In this section work on recovering the aspect angle of a recognized plane is reported. Since the recovered aspect angle of a misrecognized plane is not important, the misrecognized planes are ignored. For aspect angle recovery 91 output cells are used corresponding to each of the 7 planes available from the SRCRC simulator at the 13 aspect angles trained upon (0,5,10,...,85,90°). This makes the resultant network more complicated, but is not a major problem since SC-net networks are sparsely connected. Training and testing are done with data encoding algorithm B, a bin size of 5 and k=9 (9 generated patterns) and with a variable number of centered patterns based on thresholds of .10 and .11 as described earlier.

The recovered or predicted aspect angle is calculated by first determining to which plane the k generated patterns will be assigned. Then those $n \leq k$ patterns each have an aspect angle associated with them. The aspect angles are summed and divided by n to provide the recovered aspect angle. An example is shown in Figure 6.

Table 7 shows the results as measured by four quantities:

1. CAAED - The combined average angle error difference. It is calculated by summing the differences between the known and recovered aspect angles for all planes at all aspect angles and dividing by the number of total examples (637) minus the incorrectly classified or missed examples.

2. TAAED - The test average angle error difference. It is calculated as the sum of the difference between the known and recovered aspect angles on the **test** set only, divided by the number of test examples (504) minus the incorrectly classified or missed examples.

3. TNNTA - The total number of nearest trained angles. This measures how close the recovered angle is to the trained angle nearest the known aspect angle. For example, aspect angles at 0, 1 and 2 degrees would be nearest 0° and aspect angles at 3, 4 and 5 degrees would be nearest 5°. Hence a recovered aspect angle for a plane at 1 degree known aspect angle would have to be 0, 1 or 2 degrees to increase this count. It is a binary count of the number of test examples that have recovered aspect angles closest to the train angle nearest the known aspect angle. It can be a harsh and imprecise measure. For example a recovered angle at 12 degrees is counted as incorrect for an actual aspect angle of 13 degrees for a given plane.

4. NTAR - The nearest trained angle ratio. TNNTA over the correctly classified planes in the test set.

The total number of patterns generated, bin size and misses are also shown in Table 7. The average error in aspect angle over all correctly classified returns is less than 2 degrees and just over 2 degrees when only the test set of examples is considered. Hence, it appears that the classification process does relate the returns to the aspect angle and a prediction within 5° seems a reasonable possibility.

6 Summary

The data from real radar returns is very limited in this study. While it is too little in number to draw conclusions from, the results appear to support our findings with the SRCRCS radar data. That is, with proper data encoding it is possible to recognize airplanes from radar returns using a connectionist model.

Table 7: Results on aspect angle recovery.

Bin Size	Misses	Pattern count	CAAED	TAAED	TNNTA	NTAR
5	36	5733	1.55	2.16	292	.63
5	38	6009	1.52	2.12	292	.63
5	50	5687	1.64	2.37	284	.63

The connectionist model needs to be capable of determining it's own structure. The proper structure to enable a backpropagation network to converge was difficult to determine and convergence was slow. On the other hand, SC-net is instance-based and performs well. Cascade Correlation grows its own structure and can do reasonably, but is correct in its classification about 15% less than SC-net. Cascade Correlation was not tried with the use of fuzzy values for ranges in the input as done in SC-net. A similar encoding of its inputs might improve its performance.

Feature extraction for this problem is not obvious, in the sense that its unclear of what the best features consist. It is clear that the method of data-encoding to generate the features for train and test patterns is very important. Methods that take into account the highest peaks in the return appear best suited, such as our A.1 and B algorithms. The use of binning to group points in the return seems to be an effective method to reduce the number of input features for a learning algorithm. In fact grouping the return points into bins has a positive effect on recognition provided the bins are not made too large. As the number of patterns generated goes up, smaller bin sizes appear necessary for good discrimination among the radar returns.

When a plane is correctly recognized our results suggest that it is possible to estimate the aspect angle of the plane fairly reliably. Our method requires the use of outputs for each plane at each aspect angle trained upon, which could cause problems for some connectionist

networks that rely on a more fully connected structure than SC-net does.

It should be noted that in all of our experiments we used a ratio of train patterns to test patterns less than 1 to 3.5. Many machine learning studies use a ratio of 2 to 1 (train vs. test) [24]. Hence, the generalization properties of the network from the data are quite good. The recognition system, we described, should increase its performance with more training examples. Increasing the training examples will, for each new example, generate several (up to 9 in this study) new patterns. This will complicate the learning process by making it longer and require a larger network in the case of SC-net. However, learning time for a truly accurate system is not a major issue as learning is done just once. More training patterns will not slow the actual recognition process which is primarily bounded in time by the number of patterns presented to the network to classify a return. The training set was kept small since performance was good and to show even better performance on generated data wouldn't guarantee the same performance on a large amount of data from real radar returns.

What does this study mean for the recognition of real planes from real radar returns? It appears possible to get high recognition rates from radar returns provided the radar is good, atmospheric conditions do not disrupt the relative magnitude of the points in the return to a large degree, the encoding scheme is chosen well and changes in elevation angle do not have a tremendous effect on recognition. Changes in elevation angle have only been looked at on a couple of planes due to problems getting the data from WPAFB. The experiments we tried indicated that training patterns at different elevations for each plane will need to be included in the train set.

The recognition method described here needs to be tried on a significant set of real radar returns from real planes. Unfortunately, we do not have access to such data. The learning algorithm can be any that grows its own architecture. Its choice will have a strong affect on accuracy measurements, of course. SC-net has worked quite well in this study and should work well for real data also. Potential problems lie with the number of training patterns

needed. This will depend upon the number of returns used and the data encoding scheme. As this number gets large, most current learning schemes will become slow and may have difficulty processing the large amount of data effectively. The research reported here provides a promising guideline for future work in the recognition of airplanes from radar returns.

References

[1] Ash, T. Dynamic Node Creation in Backpropagation Networks, Tech. Report ICS 8901, Institute for Cognitive Science, UCSD, La Jolla, Ca.

[2] R.G. Atkins, R.T. Shin, J.A. Kong, A Neural Net Method for High Range Resolution Target Classification, Report from Dept. Of EE and CS and Research Lab. of Electronics, MIT, Cambridge, MA., 1988.

[3] Cherkassky, V., Lee, Y. and Lari-Najifi, Self-organizing network for regression: efficient implementation and comparative evaluation, Proc. IJCNN, Seattle, Wa. V.1, pp. 79-84, 1991.

[4] Cherkassky, V. and Lari-Najifi, Informative data representation for neural network-based diagnostic systems, Proc. 2nd Government Neural Network Applications Workshop, Huntsville, AL, 1991.

[5] S.E. Fahlman, "Faster-Learning Variations on Back-Propagation: An Empirical Study", Proceedings of 1988 Connectionist Summer School, pp.38-51, 1988.

[6] Fahlman, S. and Lebeire, C. The Cascade Correlation Learning Architecture, Carnegie Mellon, Computer Science TR, 1990.

[7] Farhout, N.H. and Babri, H., Cognitive Networks for Automated Target Recognition and Autonomous System Applications, Second Government Neural Networks Applications Workshop, Huntsville, Al. 1991.

[8] R.P. Gorman and T.J. Sejnowski, "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets", Neural Networks, Vol. 1, pp. 75-89, 1988.

[9] L.O. Hall, and S.G. Romaniuk, "A Hybrid, Connectionist, Symbolic Learning System", AAAI-90, Boston, Ma. August, 783-788, 1990.

[10] L.O. Hall, S. Romaniuk, J. Leonard, and R. Mitchell, The Use of Connectionist Networks to Recognize Airplanes from Radar Returns, Artificial Neural Networks in Engineering '91, St. Louis, Mo., pp. 921-926, Nov. 1991.

[11] Hirose, Y., Yamashita, K. and Hijiya, S., Back-Propagation Algorithm Which Varies the Number of Hidden Units, Neural Networks, V. 4, pp. 61-66, 1991.

[12] R. Hecht-Nielsen, "Kolmogorov's Mapping Neural Network Existence Theorem", Proceedings of IEEE First International Conference on Neural Networks, San Diego, CA., June, 1987.

[13] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, Reading, Ma. 1990.

[14] A Study of Machine Learning Approaches for some Classification Knowledge Bases, 4th Florida AI Research Symposium 1991, April, Cocoa Beach, pp. 125-129.

[15] Y. Lamdan and H.J. Wolfson, Geometric Hashing: A General and Efficient Model-Based Recognition Scheme, Robotics Research Lab, Courant Institute of Mathematical Sciences, NYU, New York, 1989.

[16] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Machine Learning: An Artificial Intelligence Approach. Palo Alto, Ca., Tioga Publishing, 1983.

[17] G.C. Oden, "A Symbolic Superstrate for Connectionist Models", IEEE ICNN, San Diego, California, July, 1988.

[18] S.G. Romaniuk, Extracting Knowledge from a Hybrid, Symbolic, Connectionist Network, Ph.D. Dissertation, Department of Computer Science and Engineering, University of South Florida, Tampa, 1991.

[19] Romaniuk, S.G. and Hall, L.O., Fuzzy Quantifiers and Quantifying Operators in a Connectionist Expert System Development Tool, International Joint Conference on Neural Networks, Singapore, November, pp. 134-139, 1991.

[20] S.G. Romaniuk and L.O. Hall, Divide and Conquer Neural Networks, Technical Report ISL-1219-91, Dept. of CSE, USF, Tampa, Fl.

[21] D.E. Rummelhart, and J.L. McClelland, Parallel Distributed Processing, Vol. 1, The MIT Press, Cambridge, Ma., 1986.

[22] J.W. Shavlik, R.J. Mooney, and G.G. Towell, Symbolic and Neural Learning Algorithms: An Experimental Comparison, Computer Sciences Technical Report #857, University of Wisconsin-Madison, Madison, Wisc, 1989.

[23] Wasserman, P.D., Neural Computing, Theory and Practice, Van Nostrand Rheinhold, N.Y. 1990.

[24] S.M. Weiss, and Kulikowski, C., Computer Systems That Learn, Morgan Kaufmann, San Mateo, CA. 1990.

A Divide and Conquer Networks

Error reduction on individual weights is done with the use of the Quickprop [5] weight modification algorithm,

$$\Delta w(t) = \frac{S(t)}{S(t-1) - S(t)} \Delta w(t-1)$$

where S(t) and S(t-1) are the current and previous values of $\partial E/\partial w$. This provides faster convergence, though standard gradient descent, $\Delta w(t) = -\epsilon \partial E/\partial w(t) + \alpha \Delta w(t-1)$, or the delta rule $\Delta w(t) = w(t-1) + LearningRate(actualoutput - expectedoutput) * x(t)$ (where x(t) is the output associated with the incoming link) provide comparable results [13].

Weights on links from the other cells connected to a newly introduced hidden cell are the only ones trained. This makes the training always of single layer subsets or perceptrons. All cells in the network (except the inputs) are viewed as complex feature detectors.

The Divide & Conquer Network (DCN) learning algorithm consists of two major phases; the divide phase and the conquer phase. The divide and conquer phases are done individually for each output of the problem space on which training takes place. The object is to allow each cell in the network to act as a feature detector and correctly classify some of the examples in the training set. Training is done on the connections to one cell at each step. Hence, the training is essentially on a many input, one output cell or unit at each stage. This is in the spirit of perceptrons, though the sigmoid activation function and Quickprop training rule are used instead of the actual perceptron formulae. The process is the same for each output and training can be done in parallel. Intermediate cells are not shared among outputs in the version of the learning algorithm presented in this section. There is a method of sharing the intermediate cells during learning.

Figure 7 contains a summary description of the algorithm. It begins in the conquer phase discussed later. The divide phase, which works as follows. An intermediate cell will have input connections from all the inputs, the bias and any intermediate cells on preceding layers. All the examples are presented and the weights on the input connections are trained using Quickprop error reduction. After a user set number of epochs (we have found 300 to 500 to be adequate) or error reduction stops, training halts with some of the examples being correctly classified. These examples are removed from the training set. The reduced training set is augmented by adding neighboring examples. A neighbor is defined for example E as

E', if the mean square difference between its inputs (or attributes values) is minimal. Of course E' may have a different output value than that of E. The training set resulting from the addition of the neighbors is used continue learning. The reason neighbors are added is that they are close in structure to other examples in the training set and may cause errors later if they are not considered during training.

At the point in which a new cell on the current layer has been added to the network, but has not been able to reduce the training example set, the conquer phase is entered. A new cell will be added at the next level of the network and have input connections from all of the existing cells for the output currently under consideration. Training is done and either the cell can correctly classify all of the examples from the original training set in which case the output has been learned or the divide phase is entered and processing continues in the manner outlined here. Termination will occur upon an output being correctly learned by the introduction of a cell in the conquer phase or after a set limit of cells has been used up in constructing the network. The constructed network differs in several ways from the usual backpropagation network. The inputs are directly connected to every cell not on the input layer. Hidden cells are only connected to hidden cells in other layers or one output cell. That is, the hidden structures for output cells are decoupled. This allows the hidden cells to act more as feature detectors than in a classic backpropagation network. In the spirit of Cascade Correlation networks, a pool of hidden cells of arbitrary size may be trained when a new hidden cell is being added to the network. Each starts with different random weights on its input links and the cell that provides the best results is chosen to be added to the network. During training only the weights on the incoming links to a new cell (or pool of cells) are modified.

Define:

E : Set of all training examples.

C_i : Binary Concept i to be learned.

N_i : Neighbor of example E_i.

Ed : Remaining Space (after division).

- For all output concepts C_i do

  While E has not been learned do

  Let Ed = E

  Connect all cells currently used for C_i to a new cell.

  Train the new connections and remove all correctly classified examples from the training set.

  Add neighbors N_i to Ed.

  While (Ed not empty AND Improvement) do

  -Create Feature Fdiv

  -Train Feature Fdiv on Ed

  - Remove Examples from Ed that Fdiv correctly classifies. If none were correctly classified, remove a random example and train Fdiv on the reduced example set. Continue removing examples and training until at least one example is correctly recognized.

  -To remaining Examples Ed_i in Ed add also their Neighbors N_i.

  end while.

  end while.

  The last cell added has correctly learned concept C_i as its output.

Figure 7: The DCN Algorithm

Report # 56
210-11MG-135
Prof. Kevin Kirby
Report Not Publishable at This Time

RESEARCH INITIATION PROGRAM

Mini-grant

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc

FINAL REPORT

Fiber Laser Preamplifier for Laser Radar Detectors

Prepared by:	Richard E. Miers
Academic Rank:	Associate Professor
Department and	Physics
University:	Indiana University/Purdue University at Fort Wayne
Research Location:	Indiana University/Purdue University at Fort Wayne
USAF Collaborator:	Paul F. McManamon WRDC/AARI-2 Wright Patterson AFB Dayton, OH 45433

Date: December 19, 1991

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. The assistance in administrative and directional aspects of this program by Universal Energy Systems, Inc is acknowledged.

I appreciate the support provided by my laboratory focal point Dr. Paul F. McManamon and the personnel of AARI-2. The assistance of Lt. Scott McCracken and Mike Salisbury was extremely valuable to this project.

I would like to also express my appreciation for the provision of Nd-doped fibers by Professor Elias Snitzer of Rutgers University Fiber Optics Materials Research Program and Professor Ted Morse of Brown University Division of Engineering. The information and advice accompanying these fibers has been extremely useful to this project.

FIBER LASER PREAMPLIFIER FOR LASER RADAR DETECTORS

by

Richard E. Miers, Associate Professor of Physics

ABSTRACT

Nd-doped fiber laser amplifiers for incorporation into a laser radar test system were developed around two Nd-doped fibers provided by Rutgers University and Brown University. Both fibers exhibited a fluorescent band peaking at or near 1064 nm. A gain of 10 dB was measured in an amplifier incorporating the double-clad fiber provided by Rutgers University. A second amplifier incorporating a Wavelength division multiplexer (WDM) and a pigtailed laser diode has been constructed.

FIBER LASER PREAMPLIFIER FOR LASER RADAR DETECTORS

I. INTRODUCTION

The Electro-optics Division of the Avionics Laboratory at Wright-Patterson Air Force Base is involved in the development of laser radar systems. One possible method of increasing the detection of a returning laser radar signal might be to use a fiber optical laser preamplifier immediately before the photo detector. This type of amplifier shows promise as a means of increasing the signal to noise ratio of a laser radar system detector.

During the summer of 1990 I was a USAF Summer Faculty Research Associate working with the Electro-optics Division of the Avionics laboratory at Wright-Patterson AFB. The objective of that research was to study the feasibility of developing a fiber laser amplifier for use in their laser radar test system. Since the results of that study indicated that such an amplifier could be useful, a Research Initiation Proposal for the development of this type amplifier was funded by the Air Force Office of Scientific Research.

II. OBJECTIVES OF THE RESEARCH EFFORT

A Nd-doped silica fiber is a four-level laser medium. For a four-level amplifier the unsaturated single pass gain factor can be given as

$$\gamma = \frac{\sigma \tau_f P_{abs}}{h\nu_p A_p^*} \qquad (1)$$

where the gain is given by

$$G = \frac{I_{out}}{I_{in}} = e^{\gamma}. \qquad (2)$$

I_{out}/I_{in} is the ratio of amplified signal to the input signal, σ is the stimulated emission cross section for the amplified wave, τ_f is the fluorescent lifetime of the upper lasing level, $h\nu_p$ is the energy per photon of the pump light, and P_{abs}/A^* is the effective intensity of the absorbed pump light in the fiber. [1]

For a Nd-doped silica fiber assuming $\sigma = 3 \times 10^{-20}$ cm^2, $\tau_f = 4.5 \times 10^{-4}$ s, $h\nu_p$(800 nm) $= 2.48 \times 10^{-19}$ J, $A_p^* \sim 5 \times 10^{-7}$ cm^2 gives a slope efficiency of $\gamma/P_{abs} = 0.11$/mW or $G(dB)/P_{abs} = 0.47$ dB/mW. Po, et al. reported a slope efficiency of 0.437 dB/mW in a Nd-doped silica fiber when pumped at 800 nm. [2]

The objective of this research was to construct a fiber laser amplifier using optimum amplification parameters and minimum noise to be integrated into a test laser radar system at the Electro-optics Division of the Avionics Laboratory at Wright-Patterson

Chip Parameters

Number of transistors on the chip	=	100,000
Fan-out of a typical gate on the chip	=	3
Total capacitance at an output pin	=	50 pF
Fraction of on-chip gates that switch during a clock cycle	=	0.3
Density of defects on the chip	=	$5/cm^2$

NMOS Transistor Parameters

Minimum feature size	=	1 μm
Input capacitance of feature size transistor	=	2 pF
Resistance of feature-size transistor in the depletion mode	=	25,000 Ω
Resistance of feature-size transistor in the enhancement mode	=	15,000 Ω
Power supply voltage	=	2 volts
Ratio of optimum-size to feature-size transistors	=	1

Interconnection Parameters

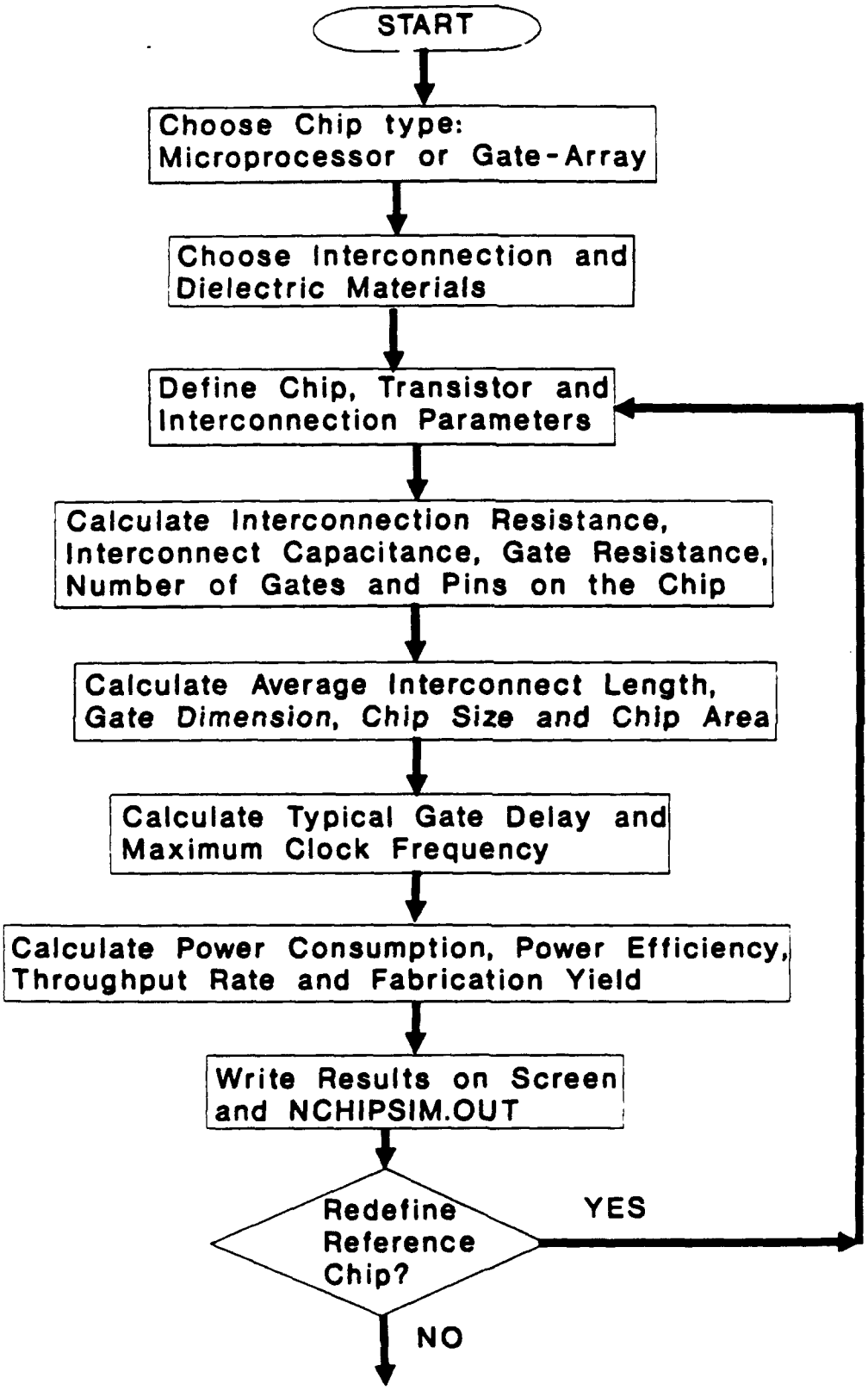
Number of interconnection layers	=	3
Widths of on-chip interconnects	=	2 μm
Pitches of on-chip interconnects	=	4 μm
Thicknesses of on-chip interconnects	=	0.2 μm

Thickness of the dielectric material	=	0.4 μm
Utilization coefficient of interconnections	=	0.33

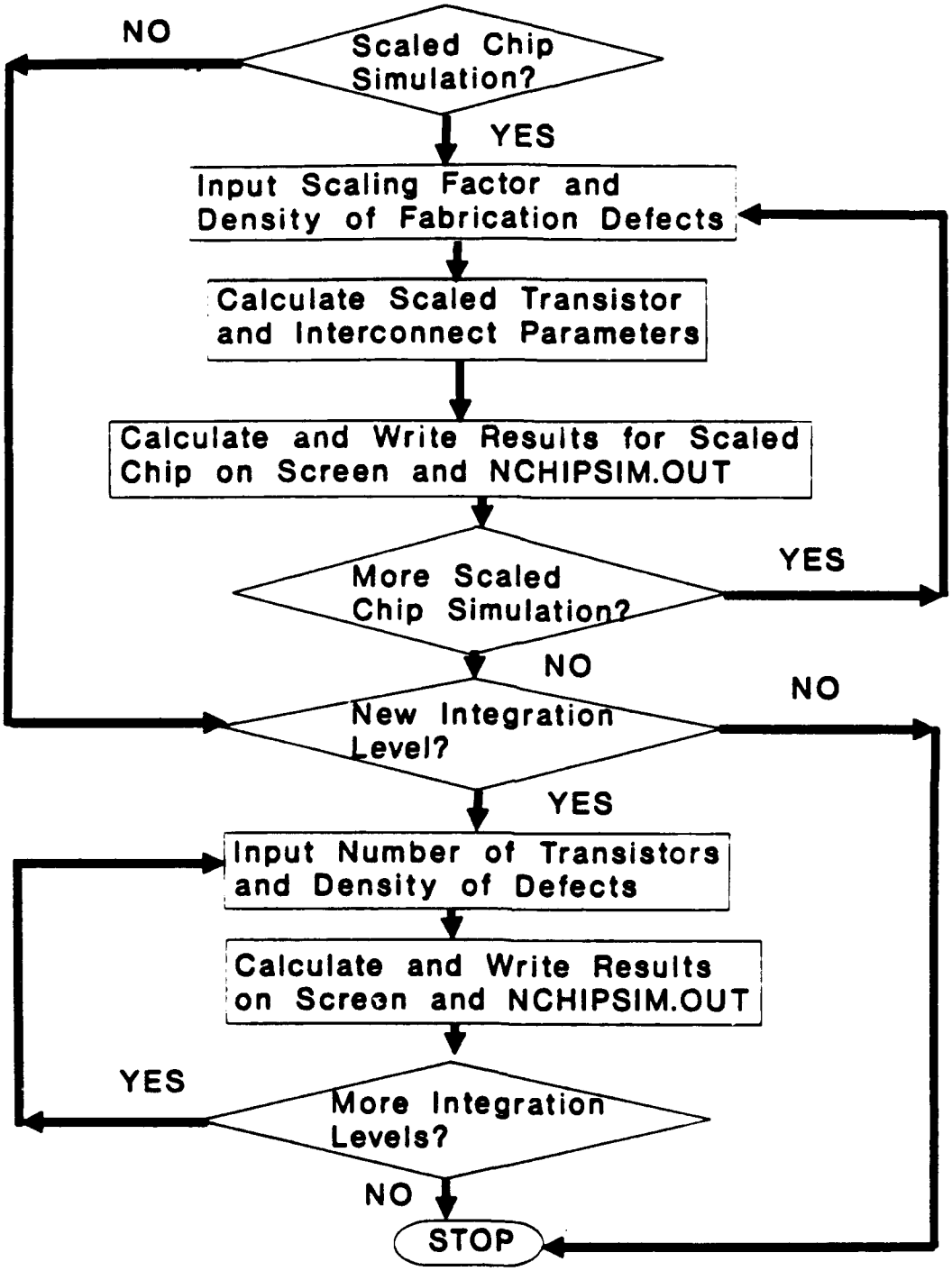
4.1 SIMULATION RESULTS USING NCHIPSIM

NCHIPSIM has been used to predict the dependence of the chip performance on its minimum feature size as well as on its integration level. For example, the dependences of the chip size, maximum clock frequency, power consumption, computational capacity, power efficiency and functional throughput rate of a 100,000-transistors silicon NMOS single-chip microprocessor on its minimum feature size in the range 0.1 to 2.5 μm are shown in figures 2 to 7, respectively and the dependences of each of these performance indicators for a 1-μm silicon NMOS microprocessor chip on its integration level in the range 100 to 100,000,000 transistors are shown in figures 8 to 13, respectively.

NCHIPSIM has also been used to compare the simulation results for several actual silicon NMOS single-chip microprocessors to the known values of the performance indicators [4-7]. For example, such a comparison for the 1.5 μm NMOS microprocessor chip called HP Focus (1982) is shown in Table 1, that for the 3 μm NMOS microprocessor chip called Stanford MIPS (1984) is shown in Table 2, that for the NMOS microprocessor chip called BERKELEY RISCI (1981) is shown in Table 3 and that for the NMOS microprocessor chip called MICRO-VAX 32720 (1984) is shown in Table 4. Tables 1-4 show that the agreement between the actual and the simulated results is indeed very good.



(Continued on Next Page)



FLOW CHART FOR "NCHIPSIM"

FIGURE 1

the typical values of the transistor related parameters for a given minimum feature size such as its input capacitance, output gate resistance, power supply voltage and the ratio of optimum-size to feature-size transistors but permits the user to change any of these values for the chip being simulated. Next, for the silicon-based chips, the user can choose a dielectric material out of silicon dioxide, polyimide, alumina and epoxy glass or select one of his own and define its dielectric constant interactively. Next, the user can choose an interconnection material out of aluminum, copper, silver, tungsten and molybdenum or select any other material and define its electrical resistivity interactively. Next, the program lists the typical values of the other interconnection parameters for the technology feature size defined earlier but allows the user to modify any of these values interactively. These parameters include the width, pitch and thickness of the on-chip interconnects, thickness of the dielectric material, number of interconnection layers and the utilization coefficient of the interconnections. Then, for the chip defined above, the program calculates and displays the values of its performance indicators such as its size, the maximum clock frequency, power consumption, computational capacity, power efficiency, functional throughput rate and the fabrication yield. Next,the program allows the user to scale the reference chip defined above by a certain scaling factor and determine the performance indicators for the scaled chip. Finally, the user is able to change the number of transistors or logic gates on the reference chip and study the dependence of the various chip performance indicators on its integration level. In addition to displaying the simulation results on the screen, the program also writes the various chip parameters and the corresponding results on an output file

for later reference. As an example, the flow chart of the program NCHIPSIM for the silicon NMOS chips is shown in figure 1. The program descriptions and the results for each of the four technologies and the details specific to a particular technology are described in the following sections.

4. THE PROGRAM "NCHIPSIM" FOR SILICON NMOS CHIPS

A microcomputer program called "NCHIPSIM" has been developed which can be used to predict the performance indicators of a microprocessor or a gate array chip based on the silicon NMOS technology as well as to study the dependence of these indicators on the feature size of the transistors and the integration level of the chip. The default values of the various chip, transistor and the interconnection parameters used in NCHIPSIM are as follows:

<u>Chip-Type Dependent Parameters for a Microprocessor Chip</u>

Interconnect-length Rent's constant	=	0.4
Pin-count Rent's constant	=	0.45
Pin-count multiplication constant	=	0.82
Logic depth	=	22

<u>Chip-Type Dependent Parameters for a Gate-Array Chip</u>

Interconnect-length Rent's constant	=	0.5
Pin-count Rent's constant	=	0.5
Pin-count multiplication constant	=	1.9
Logic depth	=	30

various component delays such as the delay due to the output resistance of the driving gate and the interconnection capacitance, the delay due to the input capacitance of the gate at the next state, the distributed-RC delay of the interconnections and the delay due to the resistance of the interconnections and the input capacitance of the gates.

e) The total delay suffered by an input signal on the chip was determined by adding the various component delays such as the delay through the logic gates including the latch delay, the combinational logic delay and the setup time, the distributed-RC delay of an interconnection that crosses the chip halfway diagonally representing the contribution of the global interconnection delay, and the contribution of the speed-of-light limit depending on the propagation speed of the electromagnetic waves on the chip. This last component was negligible unless the chip is very large or the clock frequency is greater than 1 GHz as is generally the case with the GaAs chips. The maximum clock frequency of the chip was then determined.

f) The power consumption of the chip was calculated by adding the power consumption in the logic gates and the dynamic power consumption at the I/O buffers. This depends on the power supply voltage, fraction of the on-chip gates that switch during a clock period, the total capacitance at an output pin and the number of pins per chip as determined by using the Rent's rule [2] or provided by the designer.

g) The computational capacity of the chip is a measure of the computational power of the entire chip. It was determined by using the number of gates on the chip and the maximum clock frequency of the chip determined above.

h) The power efficiency of the chip is a measure of the computational power per unit power consumption of the chip. It was obtained by using the values of the computational capacity and the power consumption determined above.

i) The functional throughput rate of a chip is a measure of the computational power per unit area of the chip. It was obtained by using the values of the computational capacity and the chip area determined above.

j) For a known value of the density of defects on the chip, the fabrication yield of the chip was obtained by using the Price law [3].

3. SOFTWARE DEVELOPMENT - GENERAL APPROACH

For each of the technologies mentioned above and using the steps outlined in the above section, flexible and user-oriented programs suitable for the simulation of the various performance indicators for the integrated circuit chips with known technology parameters were written in FORTRAN-77 and run on the IBM and its compatible personal computers. Each program was made extremely user-friendly so that it allows the user to choose the chip type, i.e., a microprocessor, a gate array or a high-speed computer chip (for CMOS and GaAs HBT technologies). Next, depending on the chip type, the program lists the constants, found empirically, required for determining the average interconnect length and the number of pins on the chip. Next, the program lists, but allows the user to change interactively, the values of several chip parameters such as the number of transistors, fan-out of a typical gate, capacitance at the output pin, density of fabrication defects and the probability of an on-chip gate to switch during a clock period. Next, the program lists

achieve higher packing densities, shorter propagation delays and smaller chips. Further, because of the much higher mobility of electrons in Gallium Arsenide (GaAs), it has emerged as a preferred substrate for the development of the very high speed integrated circuits (VHSIC). In fact, compared to the existing silicon NMOS and CMOS technologies, GaAs, MESFET, HEMT, HBT and HFET technologies have proven to be much superior for the development of the VHSIC chips.

The Device Technology Section of the Electronics Technology Laboratory at the Wright Patterson Air Force Base is interested in the development of very high speed integrated circuits based on the GaAs technology. In particular, they are concerned about the parasitic effects associated with the devices and interconnections that adversely affect the performance of a small-geometry high-speed high-density integrated circuit. They are also interested in predicting the performance of future submicrometer feature size very high speed integrated circuits. This can be accomplished by executing the following steps: a) Development of a computer-efficient model of the integrated circuit performance indicators; b) Development of a user-oriented computer program suitable for the chip performance simulation; and c) Application of the computer simulator for the determination of the performance indicators for a chip with known values of the technology parameters.

The objective of this project was to develop the computer-efficient algorithms and the related user-friendly computer software modules suitable for the simulation of the performance indicators of the submicrometer-geometry high-speed high-density integrated circuit chips based on the silicon NMOS, silicon CMOS, silicon BJT and GaAs HBT technologies. In addition to predicting the various performance indicators

of an integrated circuit chip such as its maximum clock frequency, power consumption, computational capacity, power efficiency, fabrication yield, functional throughput rate and its size, the programs are also suitable for predicting the performance of future circuits with scaled feature sizes and/or increased integration levels. In order to validate the algorithms and the related simulators, the simulation results have been compared with the data available for the existing single-chip microprocessors and gate arrays based on the NMOS and CMOS technologies.

2. ALGORITHM DEVELOPMENT - GENERAL APPROACH

a) The average interconnection length on the chip in units of the average logic gate dimension was determined by the given number of transistors or gates on the chip and the Rent's constant by using the equations derived by Donath [1].

b) For an interconnection-capacity limited VLSI chip as is the case with almost all logic intensive chips, the average logic gate dimension on the chip was obtained by setting the interconnection available per gate equal to that required per gate. This depends on the interconnection pitch (equal to the sum of the interconnection width and the spacing between the interconnects), the number of interconnection layers on the chip and the utilization coefficient of the on-chip interconnections to account for the interconnect lines not utilized or those used for power and clock distribution.

c) The value of the average interconnection length on the chip was evaluated and the values of the chip size and the chip area were determined.

d) The average delay time in a gate on the chip (defined as the time taken by the output signal to reach 50% of its steady state value) was obtained by adding the

FINAL REPORT

COMPUTER SIMULATION OF SMALL-GEOMETRY HIGH-SPEED HIGH-DENSITY INTEGRATED CIRCUIT PERFORMANCE INDICATORS

Principal Investigator: Ashok K. Goel, Ph.D.
Department of Electrical Engineering
Michigan Technological University
Houghton, MI 49931

Contract No.: F49620-88-C-0053/SB 5881-0378

Purchase Order No.: S-210-11MG-003

Date: December 5, 1991

COMPUTER SIMULATION OF SMALL-GEOMETRY HIGH-SPEED HIGH-DENSITY INTEGRATED CIRCUIT PERFORMANCE INDICATORS

ABSTRACT

For integrated circuit chips based on the silicon NMOS, silicon CMOS, silicon bipolar and GaAs heterojunction bipolar technologies, computer-efficient models of the various chip performance indicators have been developed and user-friendly computer programs called "NCHIPSIM," "CCHIPSIM," "BCHIPSIM" and "GCHIPSIM," suitable for the simulation of the chip performance indicators for a microprocessor or a gate-array chip, have been developed. In addition to predicting the various chip performance indicators such as its maximum clock frequency, power consumption, computational capacity, power efficiency, fabrication yield, functional throughput rate and the size of the chip with the given technology parameters, the programs have also been used to simulate the dependences the various chip performance indicators on the technology feature sizes in the range 0.1-5.0 microns and the chip integration levels in the range 100-1,000,000 transistors or logic gates on the chip. The results for the NMOS and CMOS chips have been compared with and found in excellent agreement with those known for several single-chip microprocessors based on these technologies.

1. INTRODUCTION

Continuous advances in the integrated circuit technology have resulted in more complex chips integrating millions of devices and interconnections. In the recent years, it has become necessary to use interconnections in two or more levels to

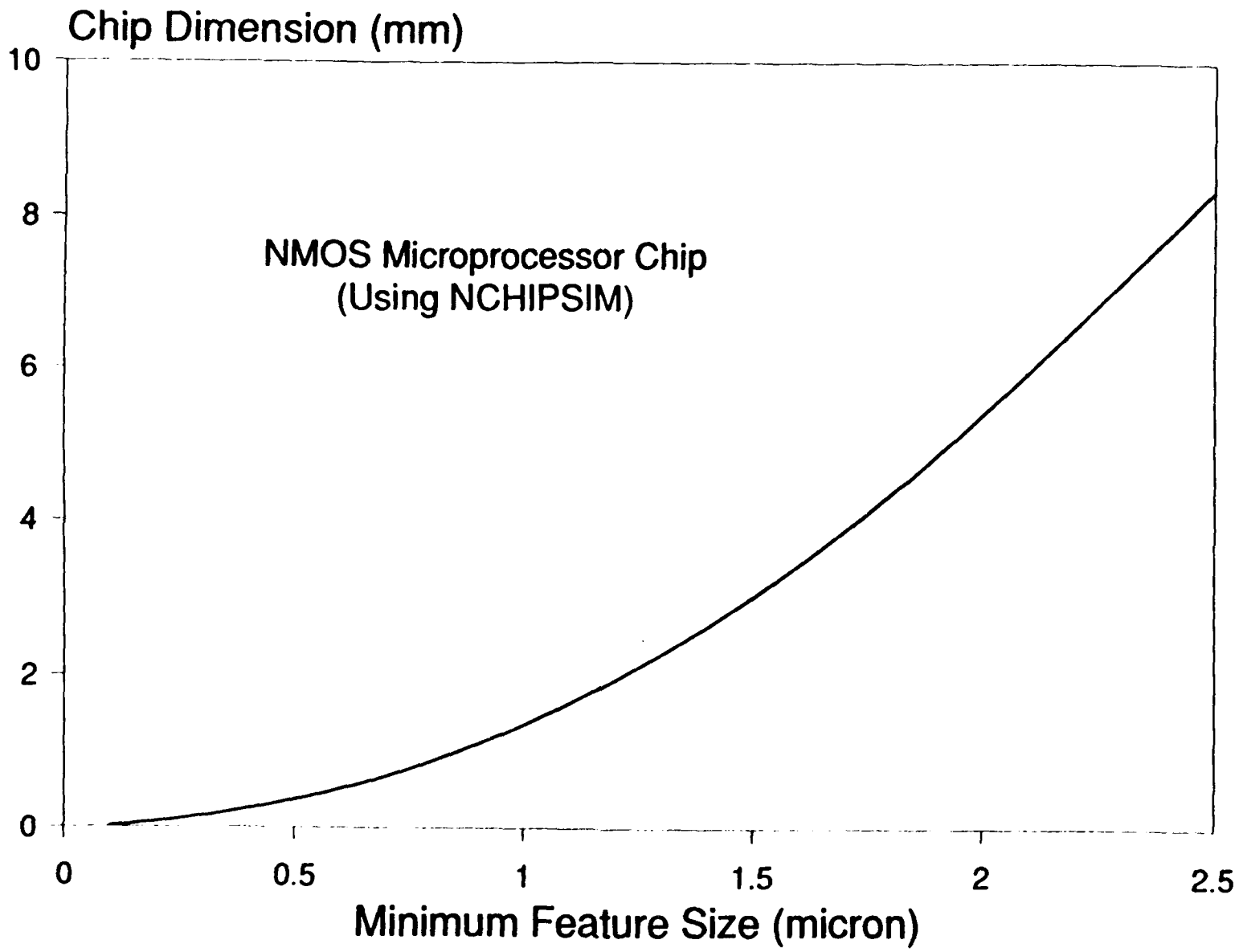

Chip Dimension (mm)
10
8
6
4
2
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 2

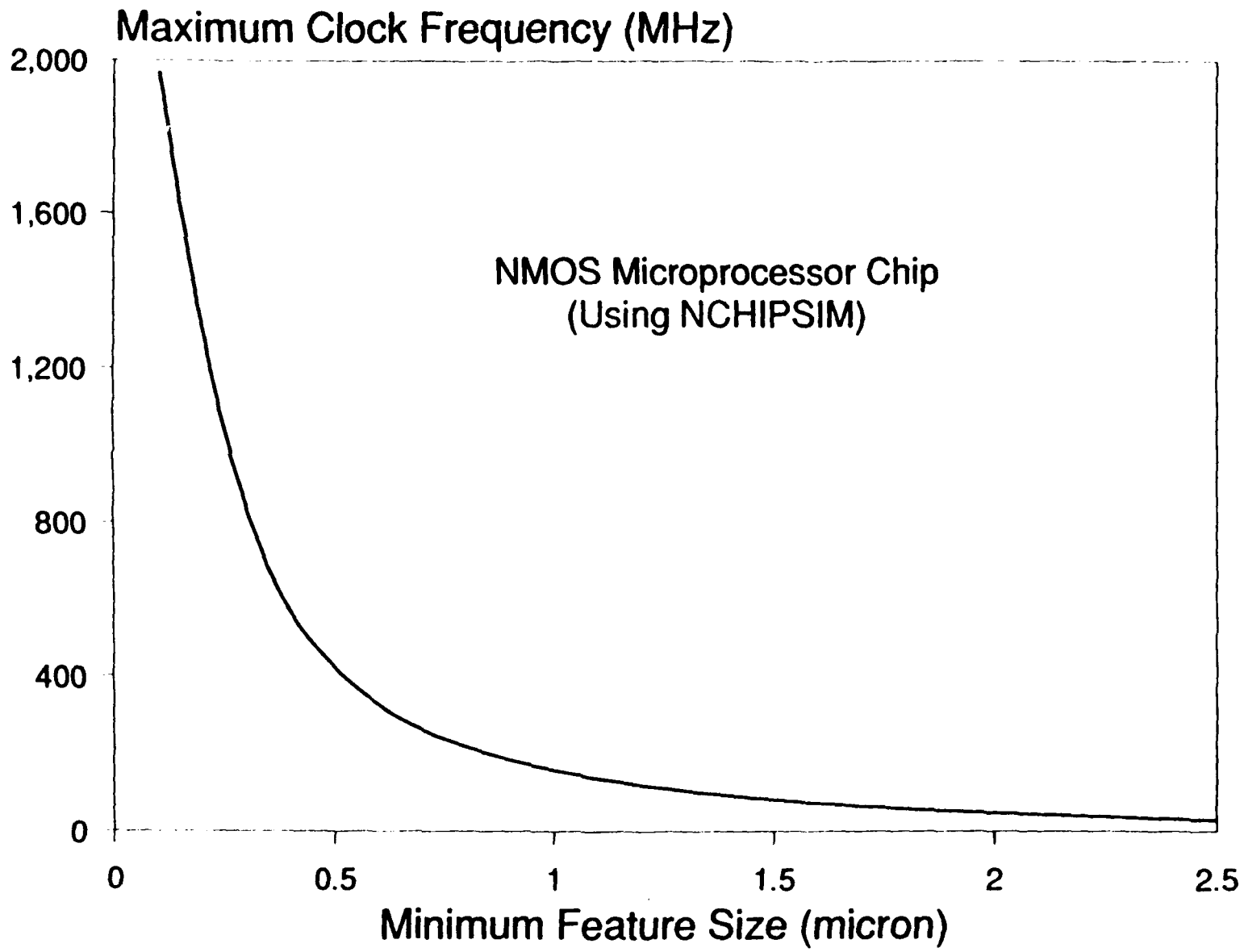

Maximum Clock Frequency (MHz)
2,000
1,600
1,200
800
400
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 3

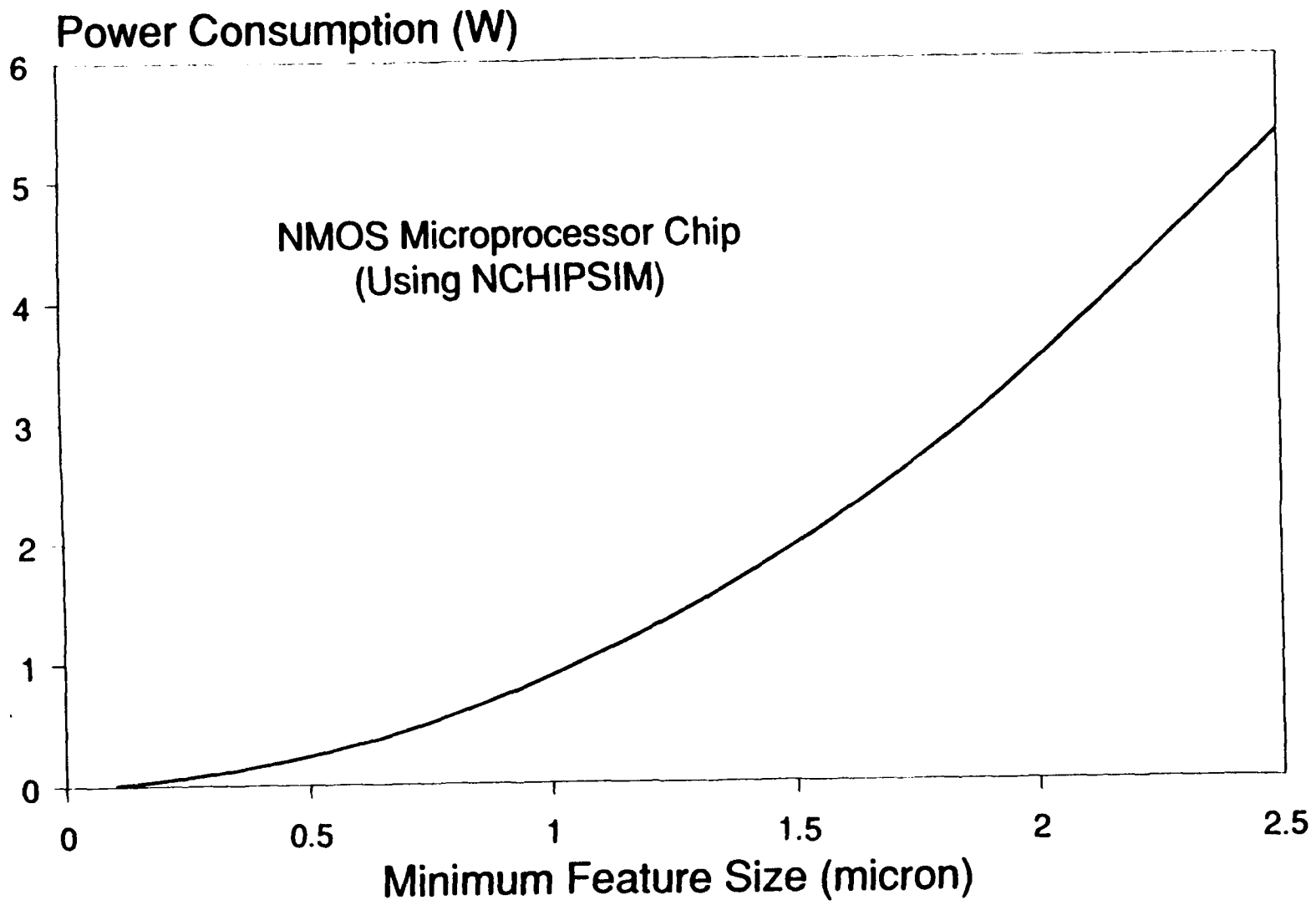

Power Consumption (W)
6
5
4
3
2
1
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 4

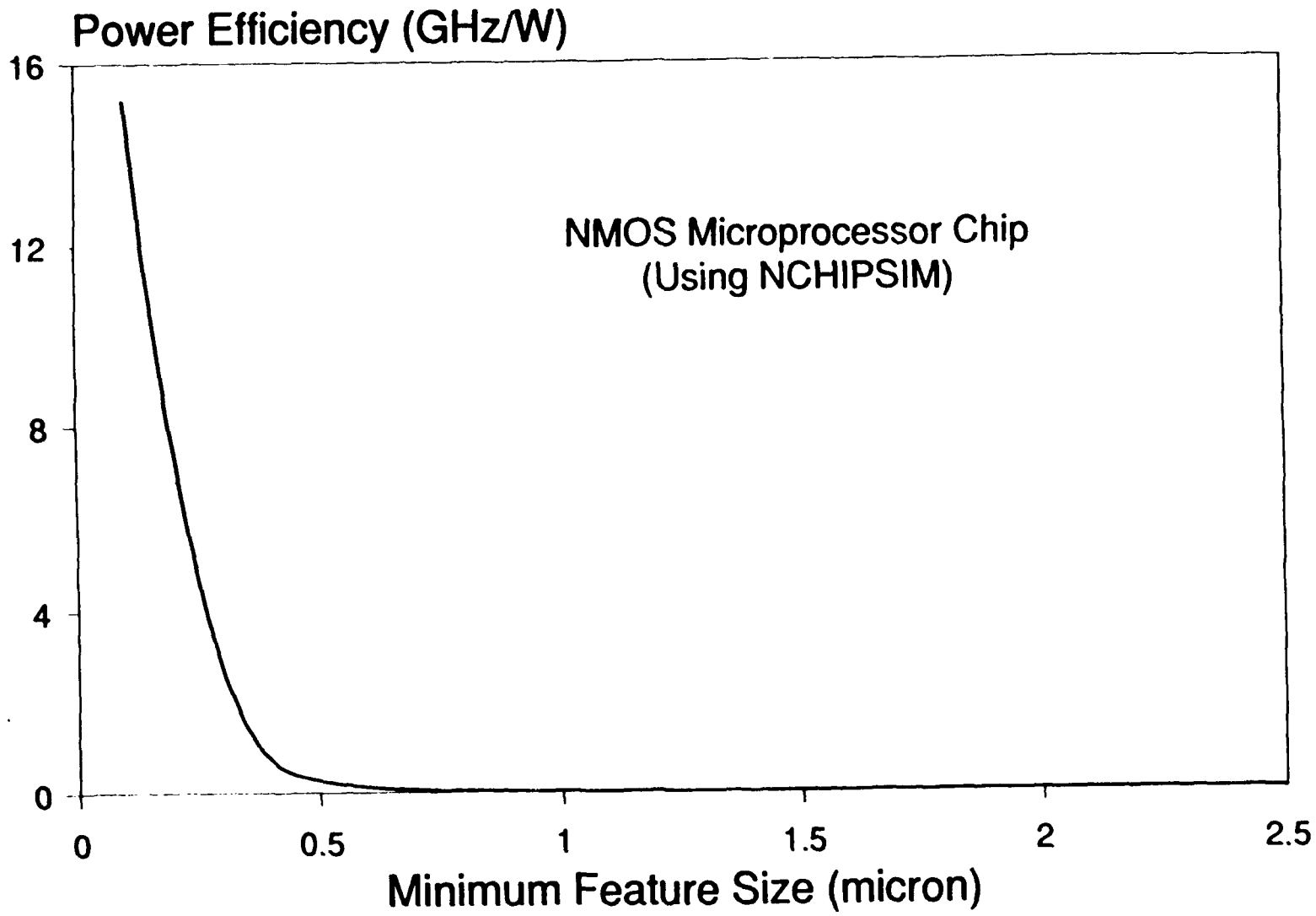

Power Efficiency (GHz/W)
16
12
8
4
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 5

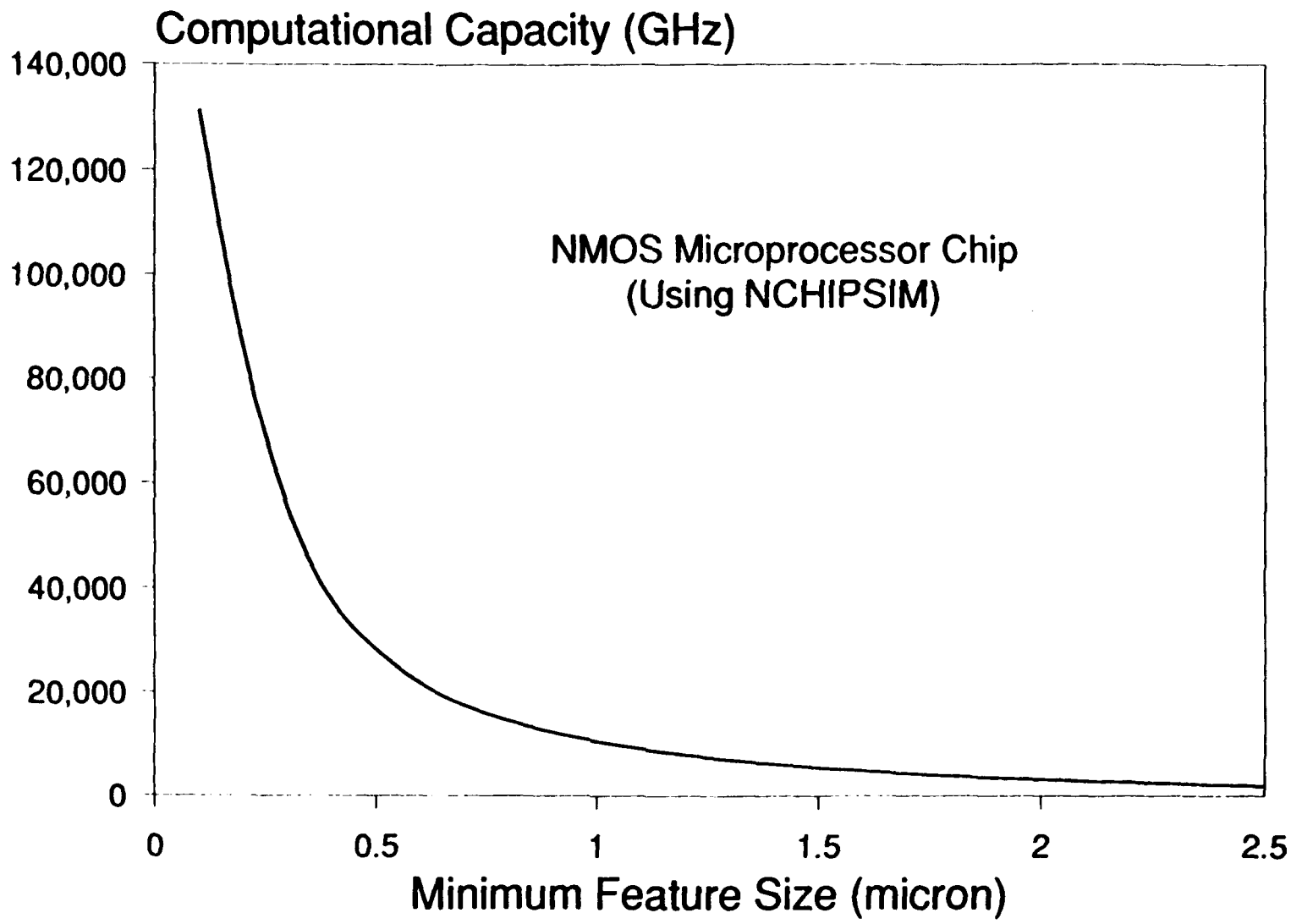

Computational Capacity (GHz)
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 6

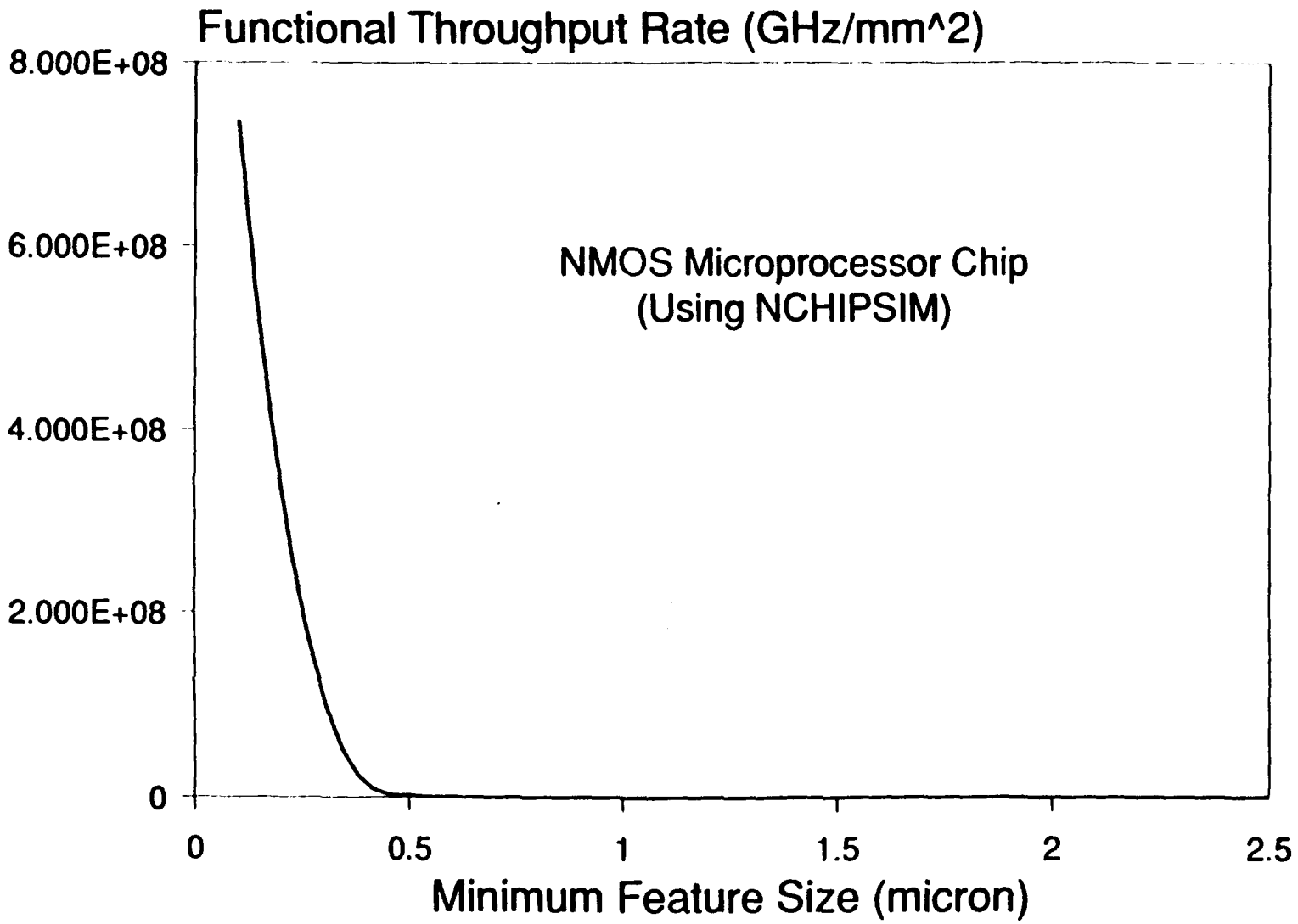

Functional Throughput Rate (GHz/mm^2)
8.000E+08
6.000E+08
4.000E+08
2.000E+08
0
NMOS Microprocessor Chip
(Using NCHIPSIM)
0
0.5
1
1.5
2
2.5
Minimum Feature Size (micron)

FIGURE 7

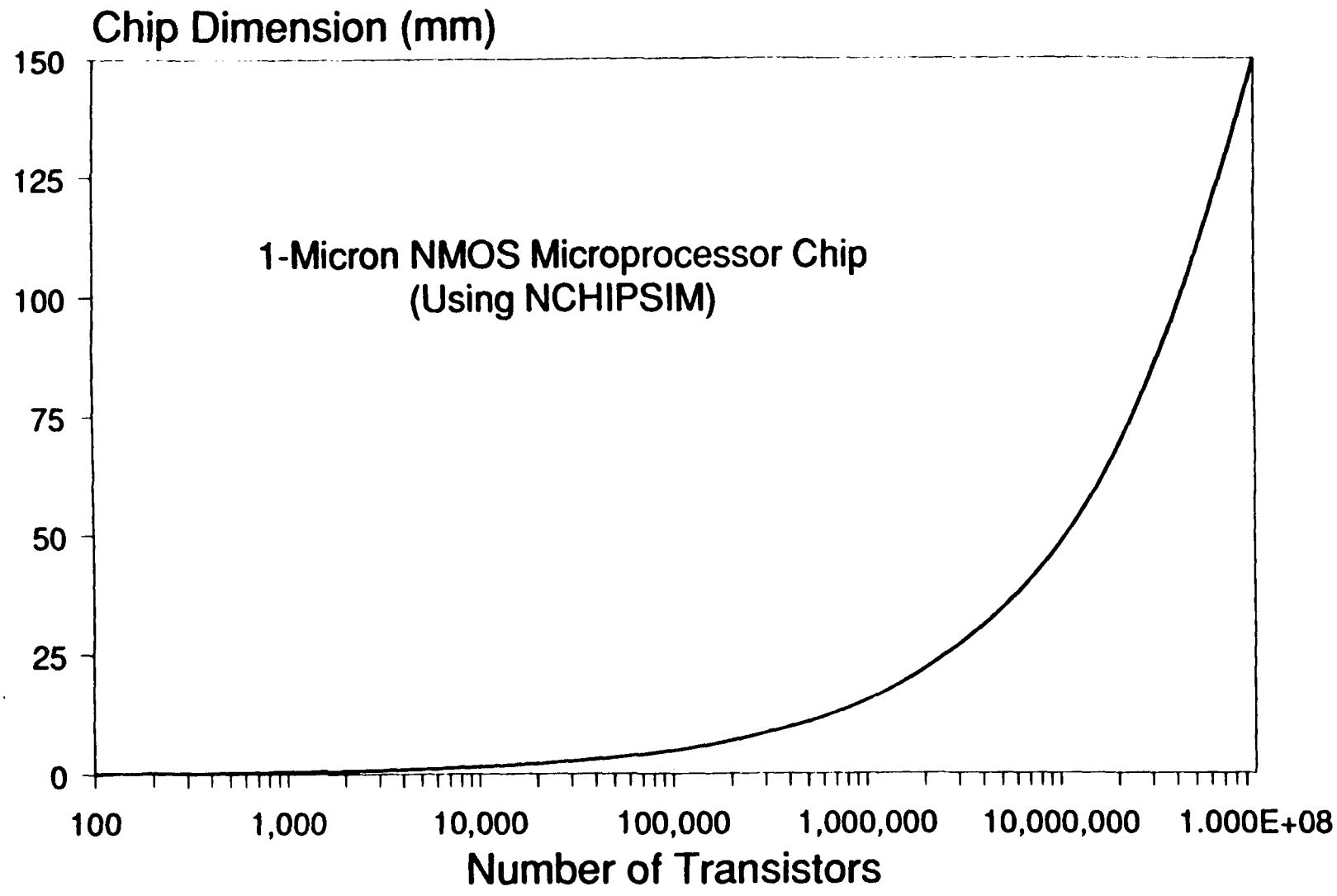

Chip Dimension (mm)
150
125
100
75
50
25
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 8

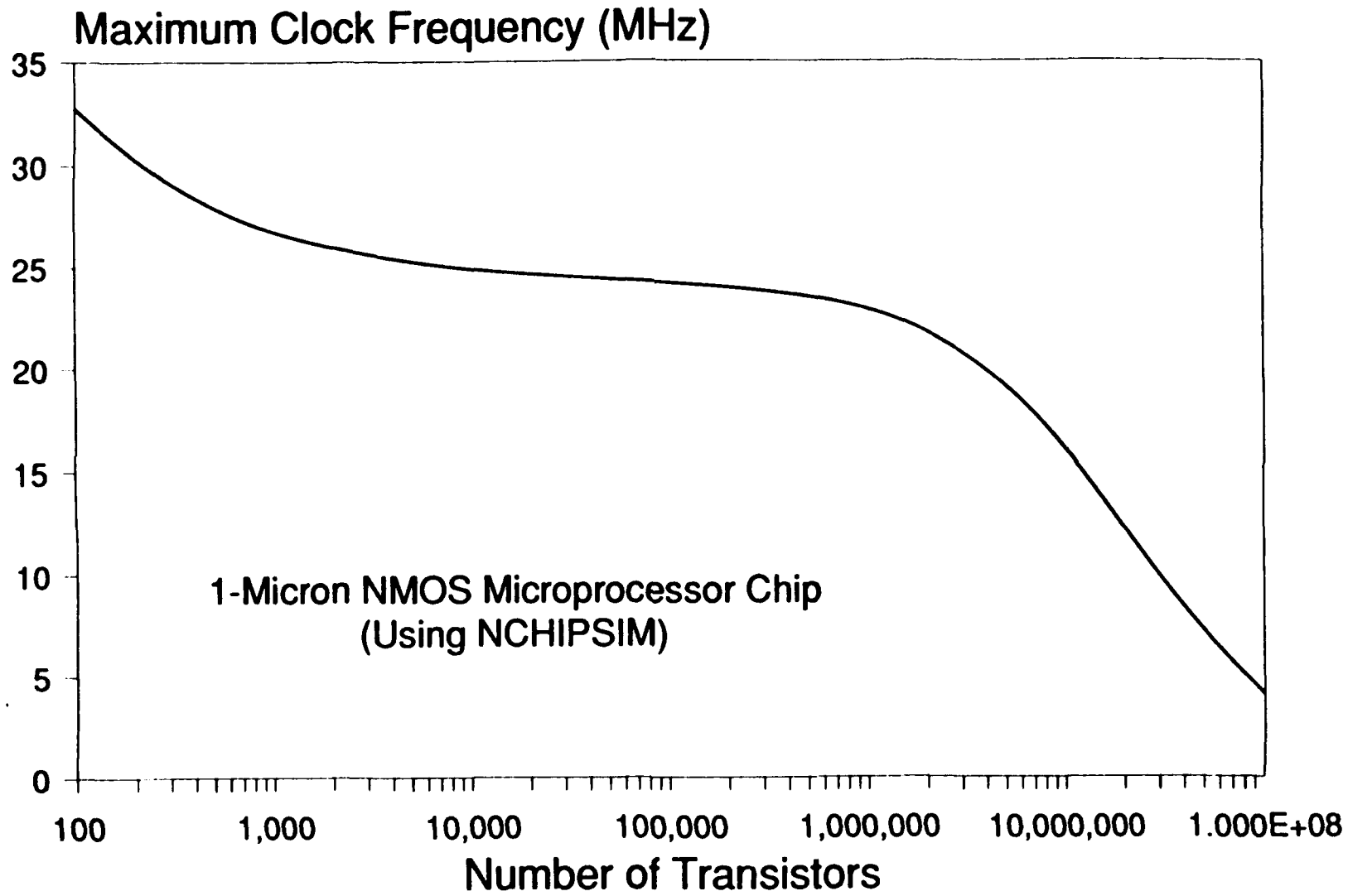

Maximum Clock Frequency (MHz)
35
30
25
20
15
10
5
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 9

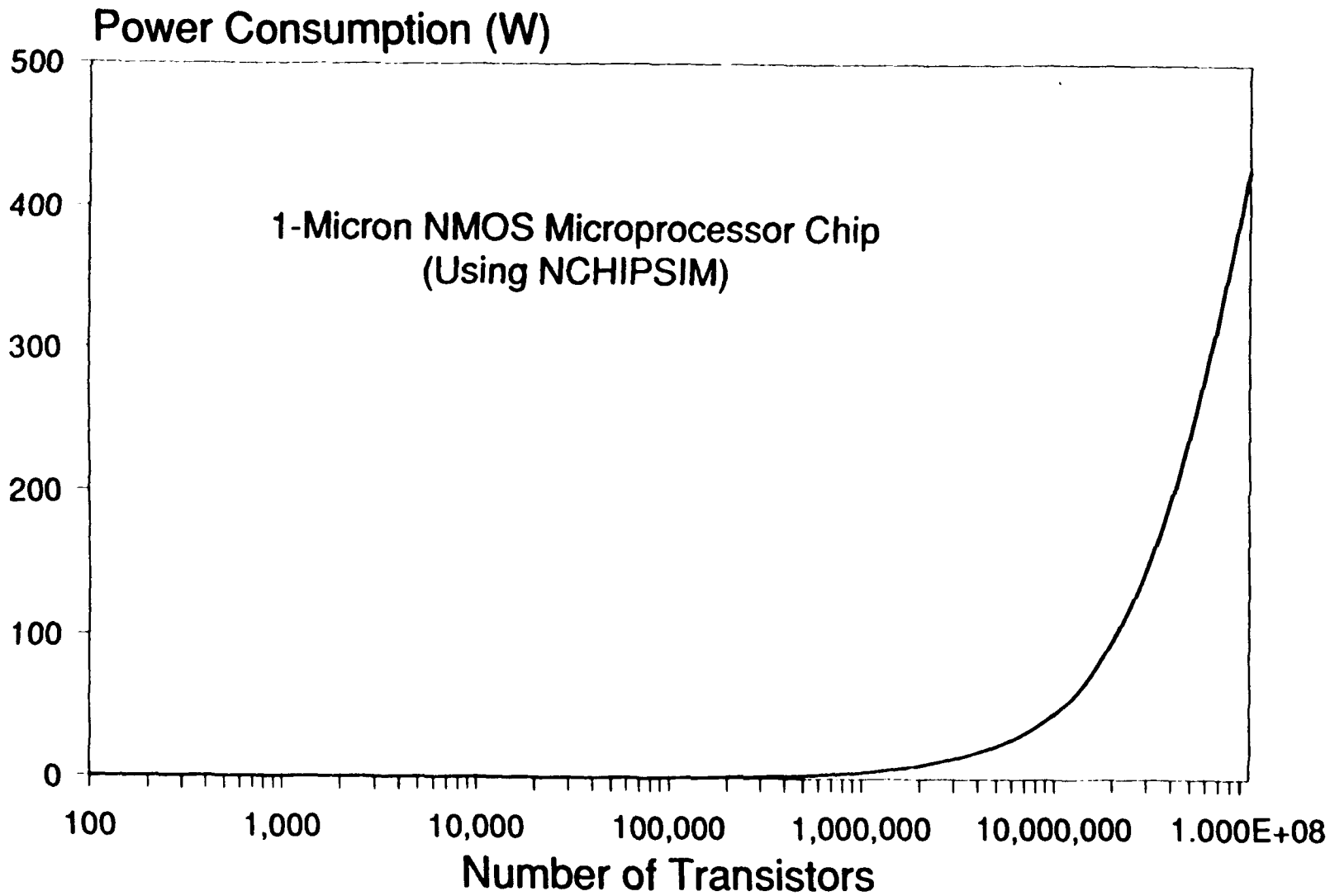

Power Consumption (W)
500
400
300
200
100
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 10

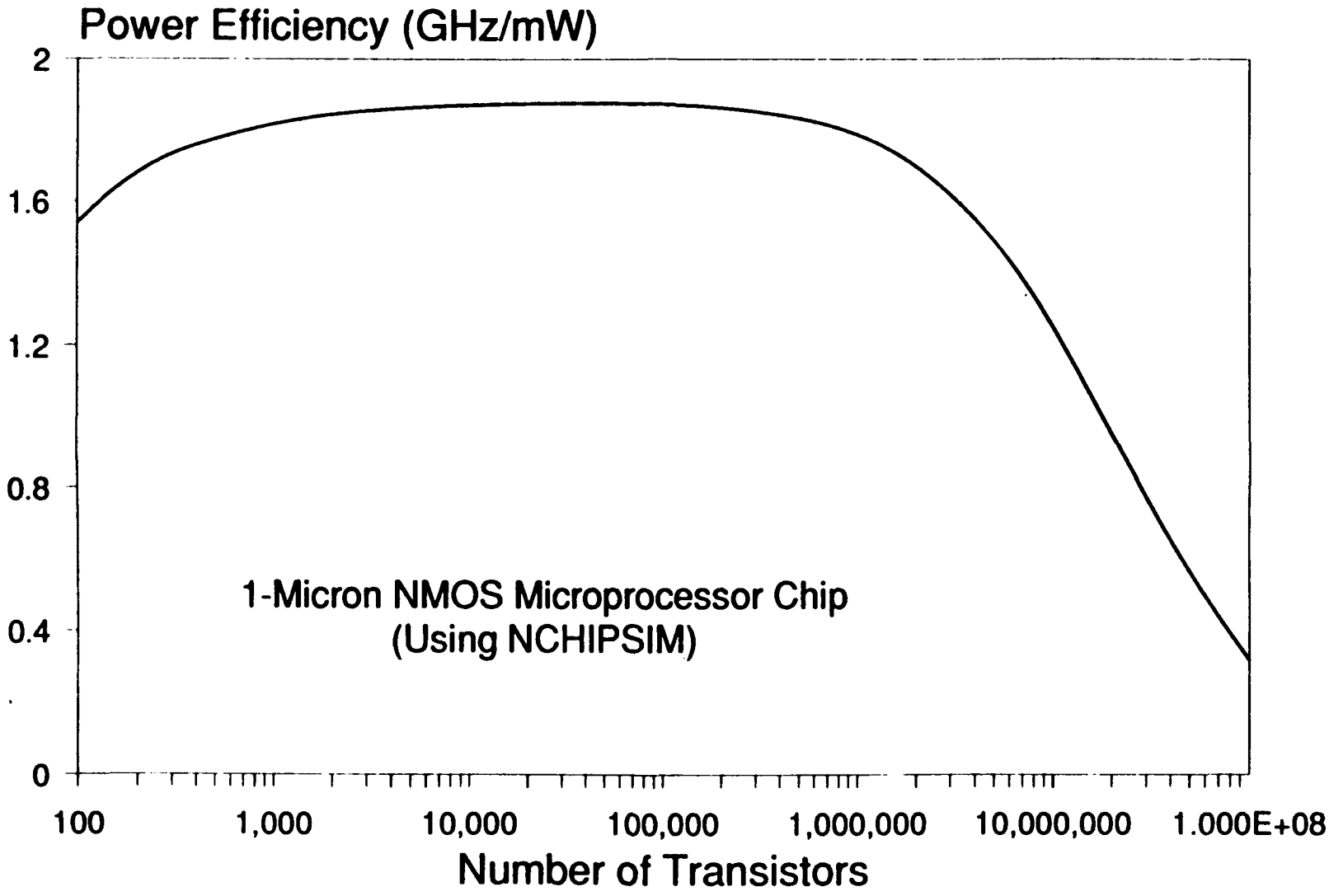

Power Efficiency (GHz/mW)
2
1.6
1.2
0.8
0.4
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 11

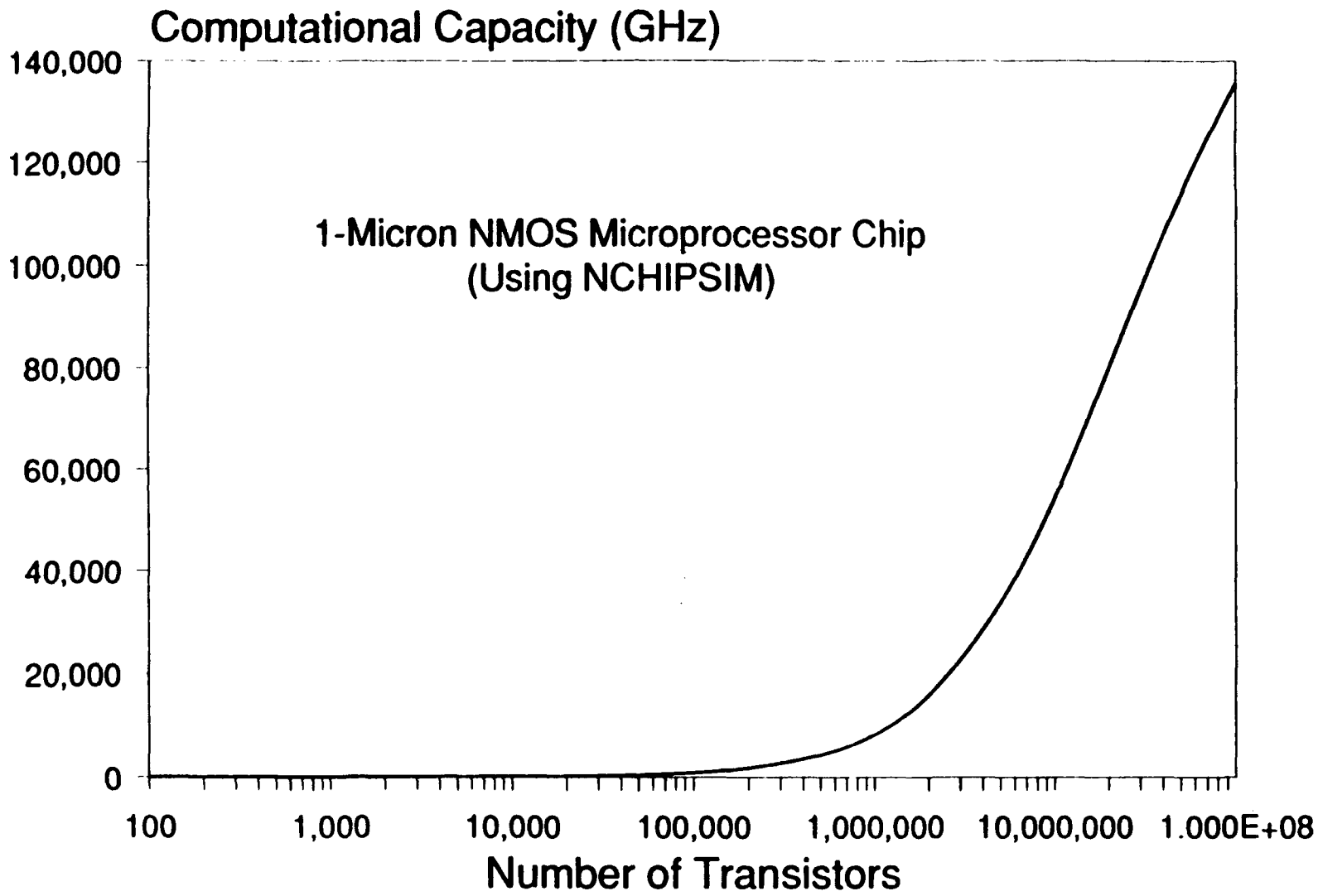

Computational Capacity (GHz)
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 12

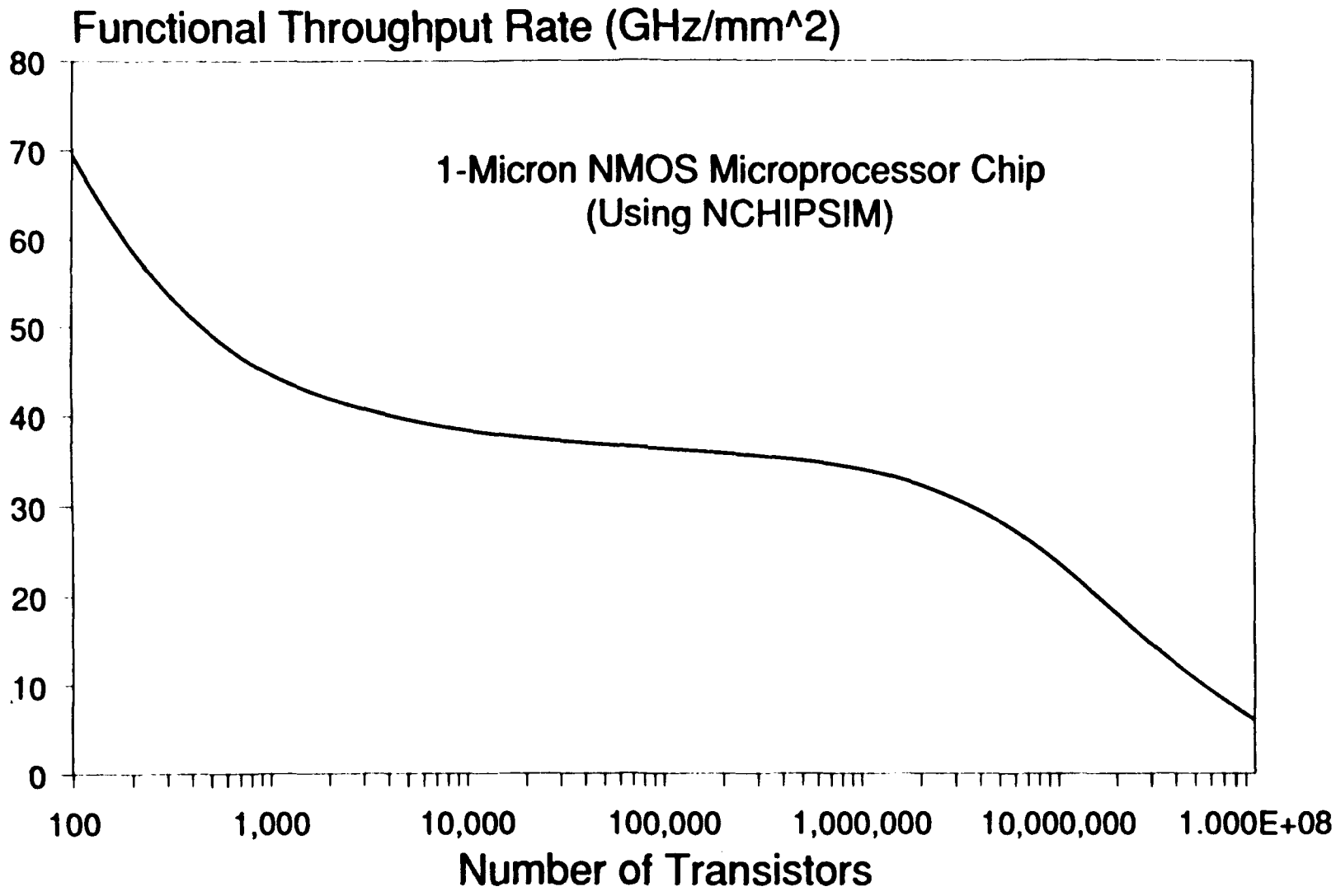

Functional Throughput Rate (GHz/mm^2)
80
70
60
50
40
30
20
10
0
1-Micron NMOS Microprocessor Chip
(Using NCHIPSIM)
100
1,000
10,000
100,000
1,000,000
10,000,000
1.000E+08
Number of Transistors

FIGURE 13

TABLE 1: Actual Data and Simulation Results for NMOS Microprocessor Chip HP FOCUS (1982)

Performance Indicator	Actual Value	Using NCHIPSIM
Chip Size	5.6	5.43 (mm)
Maximum Clock Freq.	18	22.68 (MHz)
Power Consumption	4	4.61 (Watt)
Computational Cap.	*	2551.8 (GHz)
Power Efficiency	*	0.55 (GHz/mW)
Functional Throughput	*	86.63 (GHz/mm^2)
Fabrication Yield	*	40.44 (%)

Chip Parameters	
Minimum Feature Size	= 1.5 microns
Number of Transistors	= 450,000
Interconnect Layers	= 3
Number of Pins	= 83
Interconnection Pitch	= 2.5 microns

* Data is not available

TABLE 2: Actual Data and Simulation Results for NMOS Microprocessor Chip STANFORD MIPS (1984)

Performance Indicator	Actual Value	Using NCHIPSIM
Chip Size	5.8	6.7 (mm)
Maximum Clock Freq.	4	4.63 (MHz)
Power Consumption	2	2.07 (Watt)
Computational Cap.	*	27.76 (GHz)
Power Efficiency	*	13.43 (MHz/mW)
Functional Throughput	*	619.1 (MHz/mm^2)
Fabrication Yield	*	30.85 (%)

Chip Parameters	
Minimum Feature Size	= 3.0 microns
Number of Transistors	= 24,000
Interconnect Layers	= 2
Number of Pins	= 84
Interconnection Pitch	= 9.0 microns

* Data is not available

TABLE 3: Actual Data and Simulation Results for NMOS Microprocessor Chip BERKELEY RISCI (1981)

Performance Indicator	Actual Value	Using NCHIPSIM
Chip Size	9.0	10.06 (mm)
Maximum Clock Freq.	4	4.9 (MHz)
Power Consumption	*	3.73 (Watt)
Computational Cap.	*	54.5 (GHz)
Power Efficiency	*	14.6 (MHz/mW)
Functional Throughput	*	537.8 (MHz/mm^2)
Fabrication Yield	*	16.48 (%)

Chip Parameters	
Minimum Feature Size	= 4.0 microns
Number of Transistors	= 44,500
Interconnect Layers	= 2
Number of Pins	= 54
Interconnection Pitch	= 12.0 microns

* Data is not available

TABLE 4: Actual Data and Simulation Results for NMOS Microprocessor Chip MICRO-VAX 32720 (1984)

Performance Indicator	Actual Value	Using NCHIPSIM
Chip Size	8.6	10.27 (mm)
Maximum Clock Freq.	10	10.63 (MHz)
Power Consumption	3	2.62 (Watt)
Computational Cap.	*	332.1 (GHz)
Power Efficiency	*	126.8 (GHz/mW)
Functional Throughput	*	3.15 (GHz/mm^2)
Fabrication Yield	*	15.95 (%)

Chip Parameters	
Minimum Feature Size	= 2.0 microns
Number of Transistors	= 125,000
Interconnect Layers	= 3
Number of Pins	= 68
Interconnection Pitch	= 9.0 microns

* Data is not available

5. THE PROGRAM "CCHIPSIM" FOR SILICON CMOS CHIPS

A microcomputer program called "CCHIPSIM" has been developed which can be used to predict the performance indicators of a microprocessor, a gate array or a high-speed computer chip based on the silicon CMOS technology as well as to study the dependence of these indicators on the feature size of the transistors and the integration level of the chip. The default values of the various chip, transistor and the interconnection parameters used in CCHIPSIM are as follows:

Chip-Type Dependent Parameters for a Microprocessor Chip

Interconnect-length Rent's constant	=	0.4
Pin-count Rent's constant	=	0.45
Pin-count multiplication constant	=	0.82
Logic depth	=	22

Chip-Type Dependent Parameters for a Gate-Array Chip

Interconnect-length Rent's constant	=	0.5
Pin-count Rent's constant	=	0.5
Pin-count multiplication constant	=	1.9
Logic depth	=	30

Chip-Type Dependent Parameters for a High-Speed Computer Chip

Interconnect-length Rent's constant	=	0.6
Pin-count Rent's constant	=	0.63
Pin-count multiplication constant	=	1.4
Logic depth	=	10

Chip Parameters

Number of transistors on the chip	=	120,000
Fan-out of a typical gate on the chip	=	3
Number of logic gates on the chip	=	20,000
Total capacitance at an output pin	=	50 pF
Fraction of on-chip gates that switch during a clock cycle	=	0.3
Density of defects on the chip	=	$5/cm^2$

Transistor Parameters

Minimum feature size	=	1 μm
Input capacitance of feature size NMOS transistor	=	2 fF
Power supply voltage	=	2.5 volts
Ratio of optimum-size to feature-size transistors	=	1
Ratio of W/L of PMOS to W/L of NMOS	=	2
Output resistance of a typical gate	=	30,000 Ω
Time delay in a typical gate	=	7,914 ps

Interconnection Parameters

Number of interconnection layers	=	3
Widths of on-chip interconnects	=	2 μm
Pitches of on-chip interconnects	=	4 μm
Thicknesses of on-chip interconnects	=	0.25 μm

Thickness of the dielectric material	=	0.2 μm
Utilization coefficient of interconnections	=	0.4
Interconnection resistance	=	56 Ω/mm
Interconnection capacitance	=	0.48 pF/mm

5.1 SIMULATION RESULTS USING CCHIPSIM

CCHIPSIM has been used to predict the dependence of the chip performance on its minimum feature size as well as on its integration level. For example, the dependences of the chip size, maximum clock frequency, power consumption, computational capacity, power efficiency and functional throughput rate of a 20,000-logic gates silicon CMOS single-chip microprocessor on its minimum feature size in the range 0.1 to 5 μm are shown in figures 14 to 19, respectively; and the dependences of each of these performance indicators for a 1 μm silicon CMOS microprocessor chip on its integration level in the range 100 to 1,000,000 logic gates are shown in figures 20 to 25, respectively.

CCHIPSIM has also been used to compare the simulation results for several actual silicon CMOS single-chip microprocessors to the known values of the performance indicators [4-7]. For example, such a comparison for the 2 μm CMOS microprocessor chip called Fairchild Clipper (1985) is shown in Table 5 and that for the 1.5 μm CMOS microprocessor chip called Intel 80386 (1985) is shown in Table 6. Tables 5 and 6 show that the agreement between the actual and the simulated results is again very good.

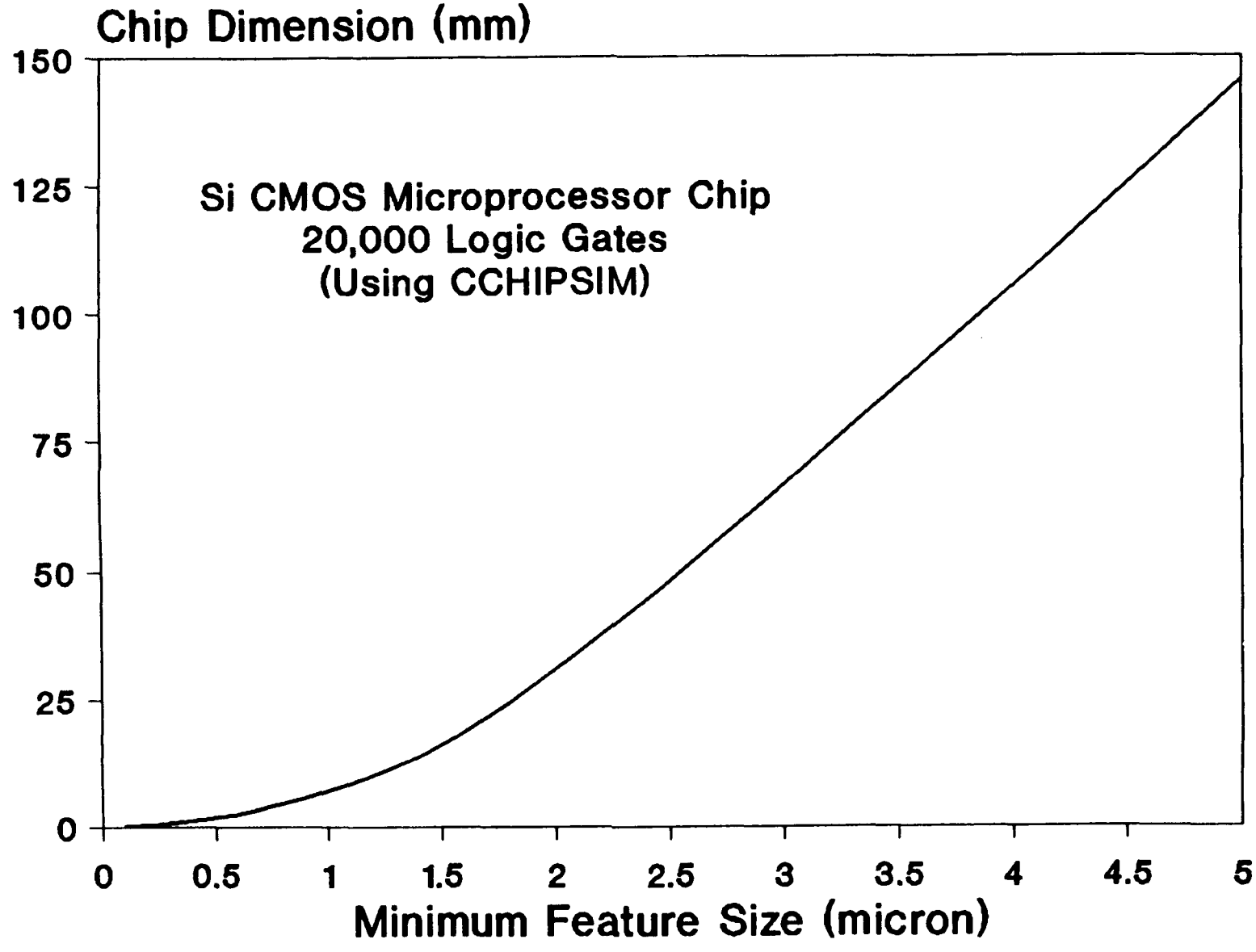

Chip Dimension (mm)
150
125
100
75
50
25
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 14

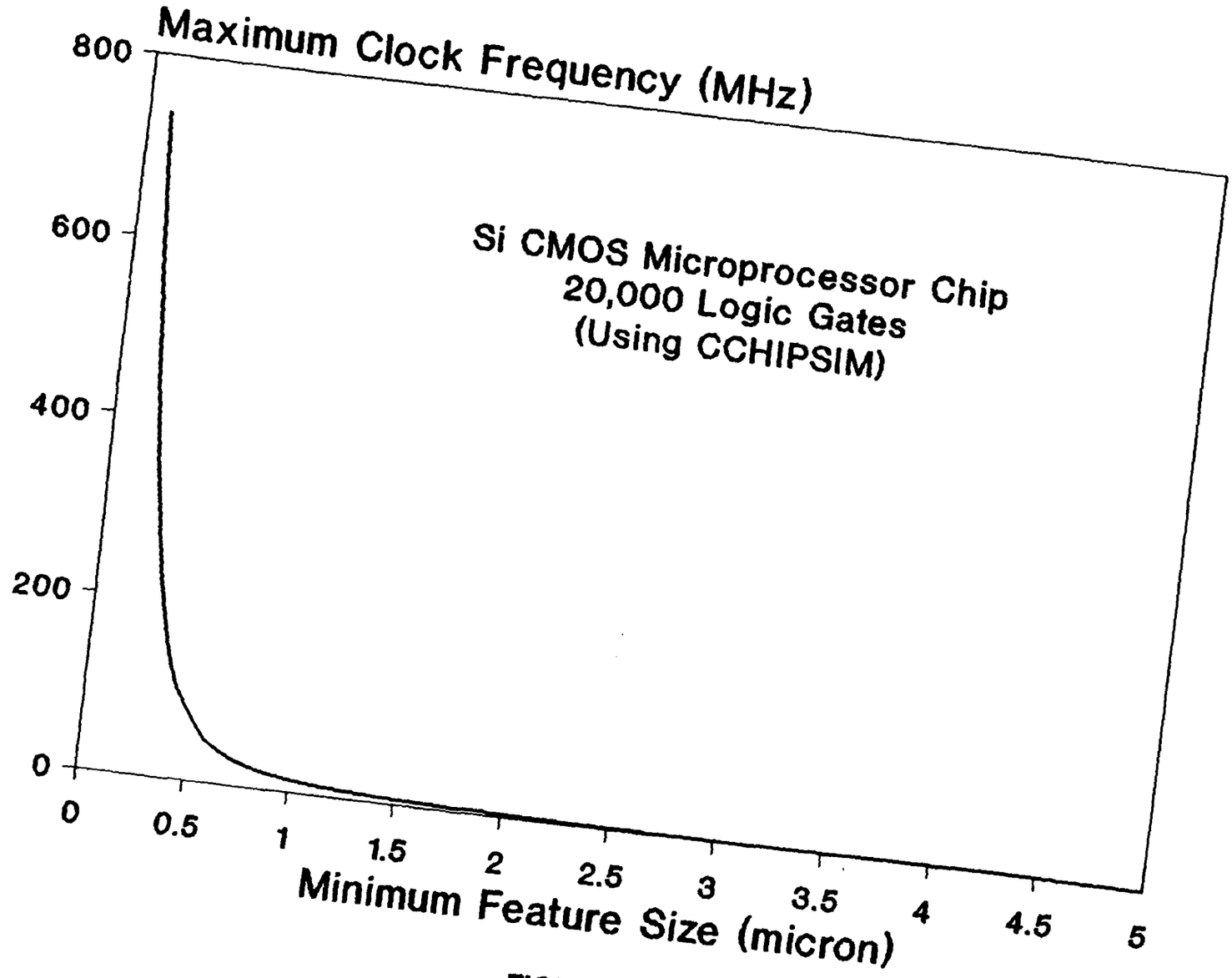

Maximum Clock Frequency (MHz)
800
600
400
200
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 15

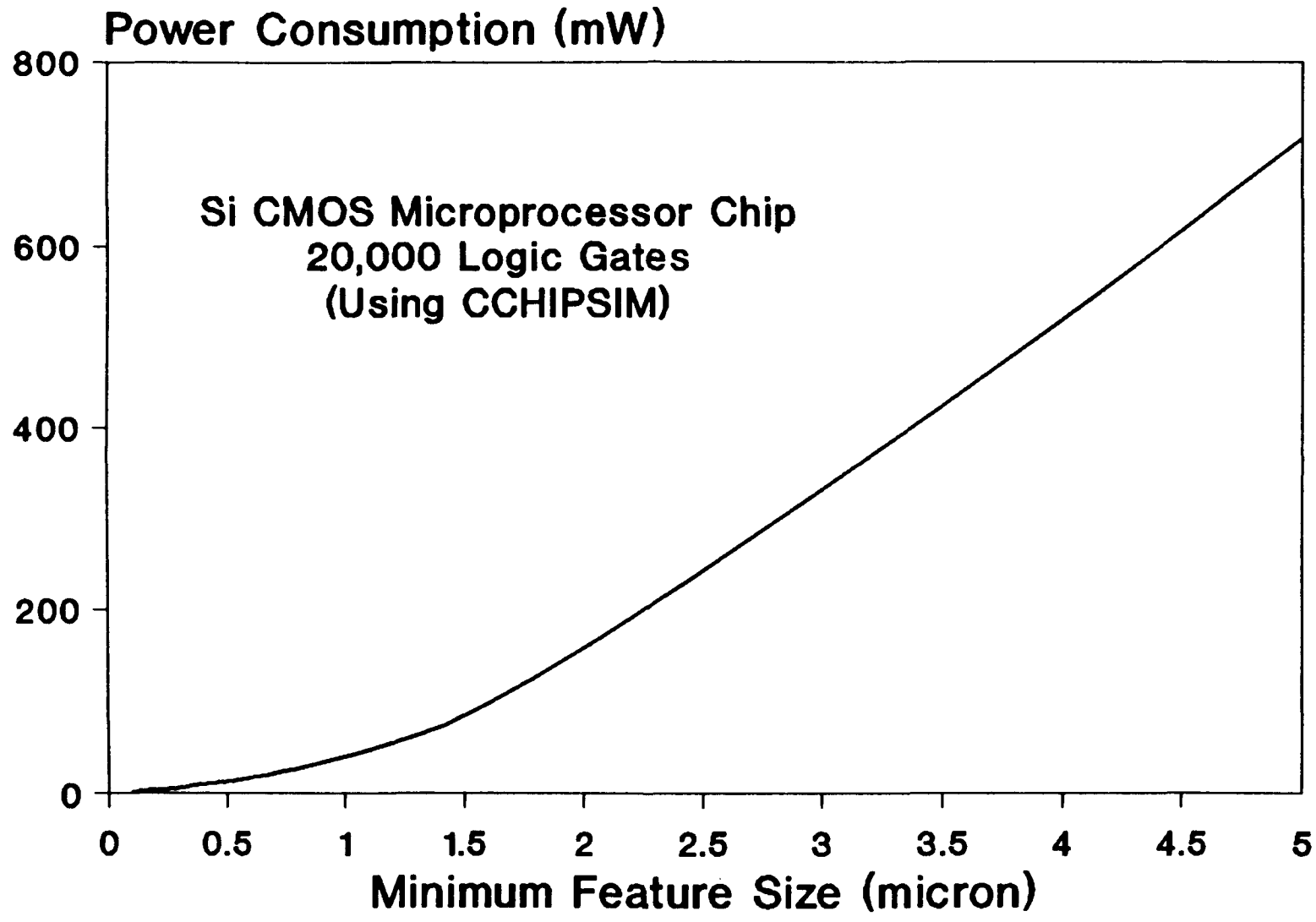

Power Consumption (mW)
800
600
400
200
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 16

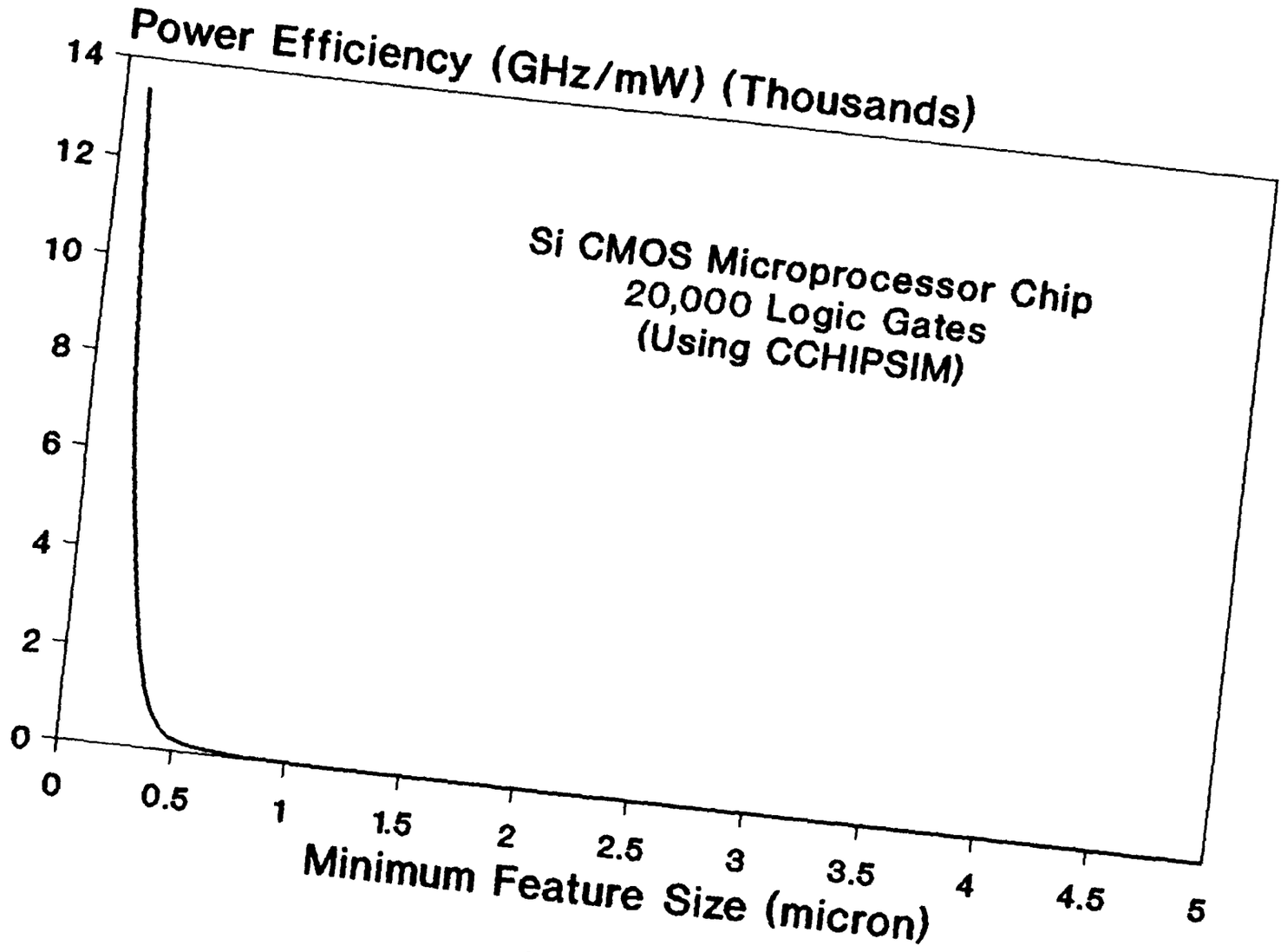

Power Efficiency (GHz/mW) (Thousands)
14
12
10
8
6
4
2
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 17

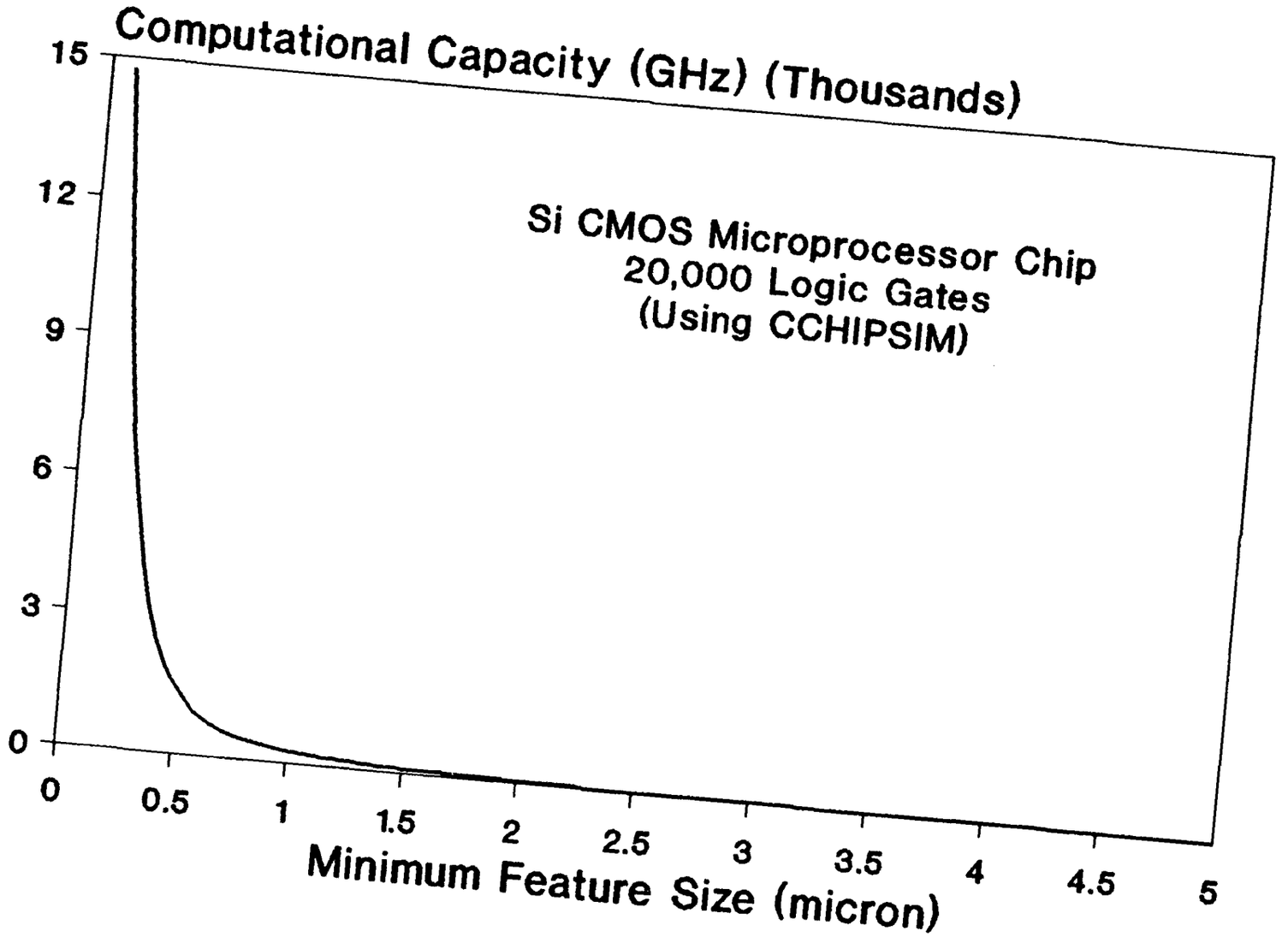

Computational Capacity (GHz) (Thousands)
15
12
9
6
3
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 18

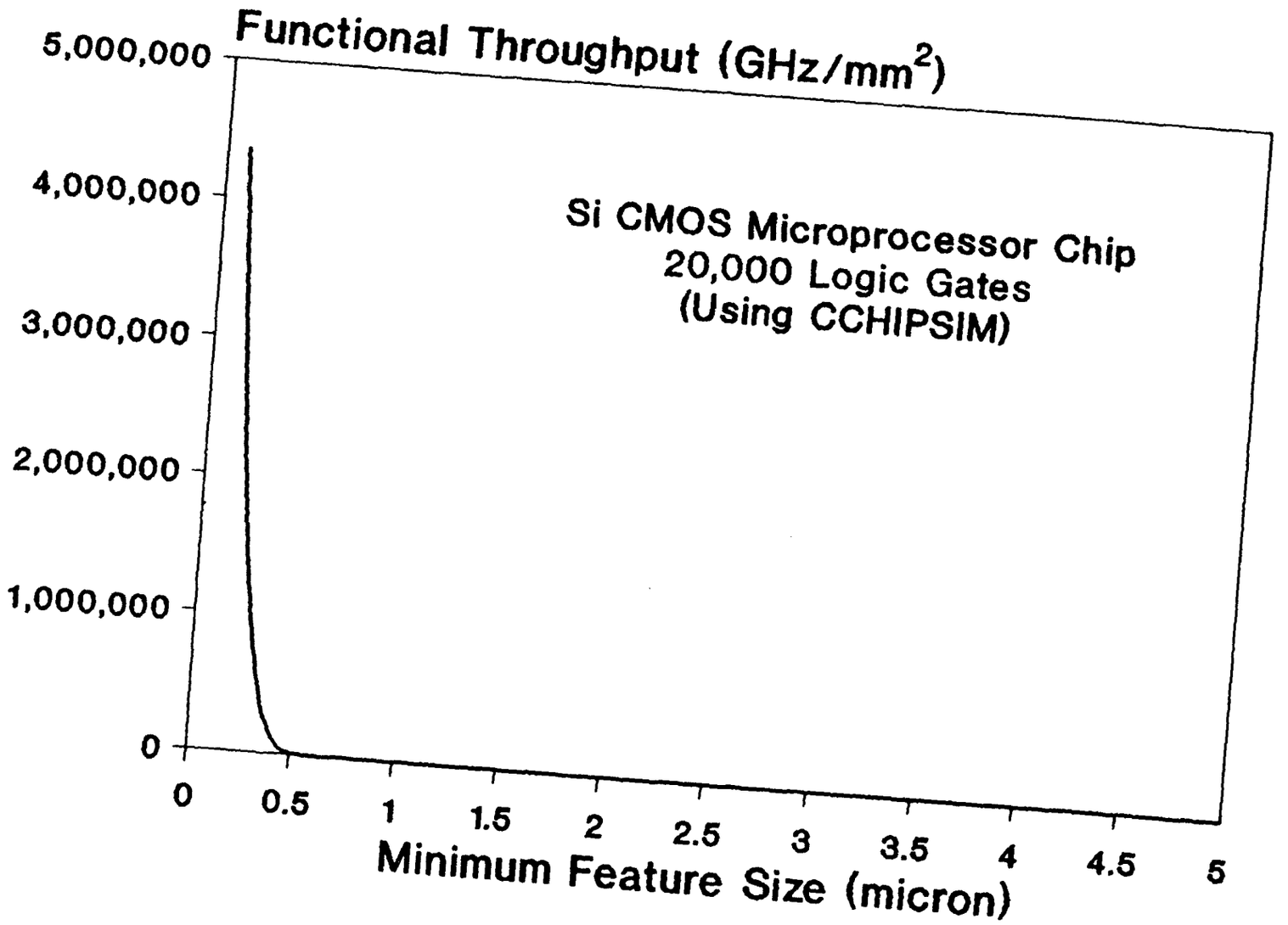

Functional Throughput (GHz/mm^2)
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
Si CMOS Microprocessor Chip
20,000 Logic Gates
(Using CCHIPSIM)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Minimum Feature Size (micron)

FIGURE 19

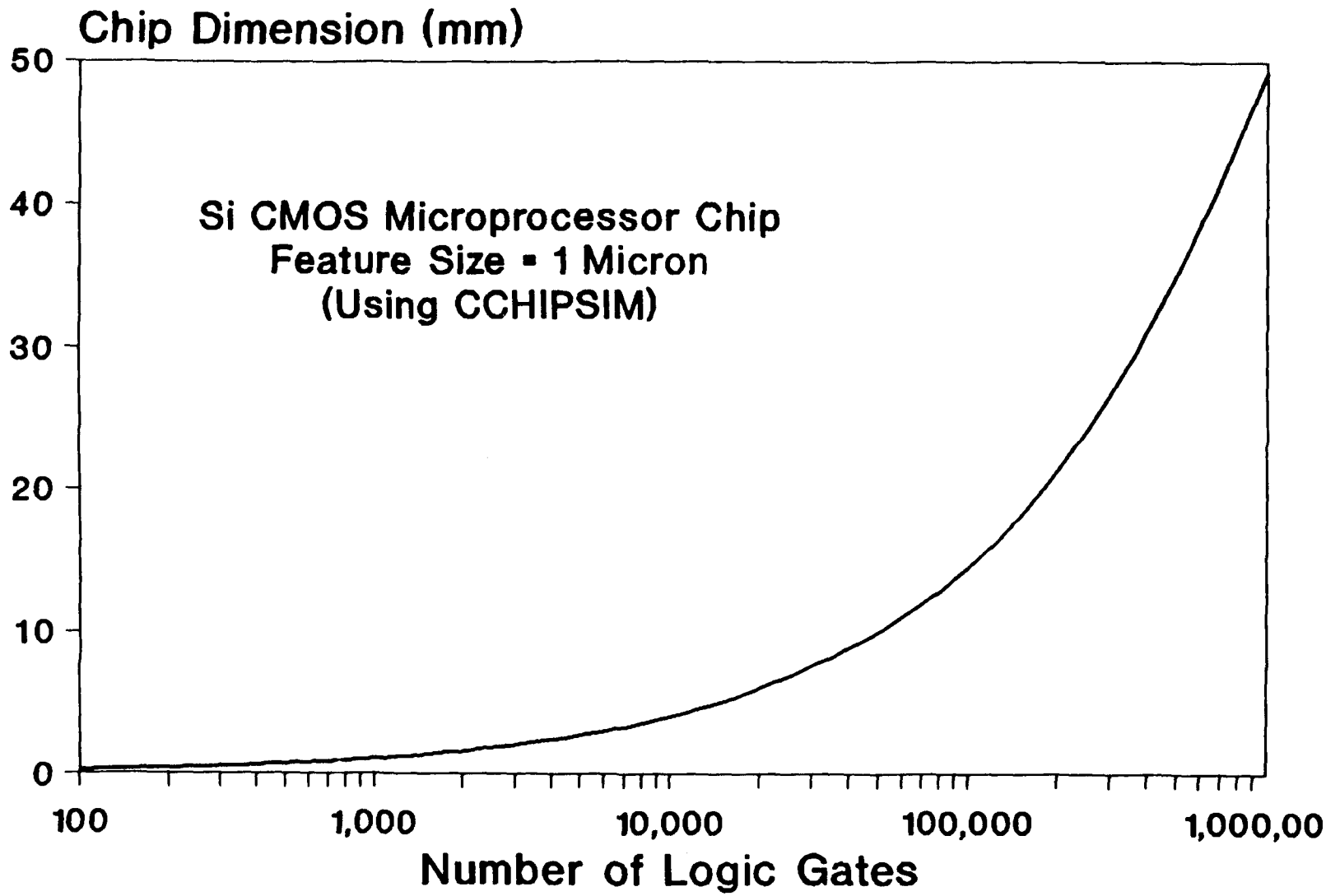

Chip Dimension (mm)
50
40
30
20
10
0
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 20

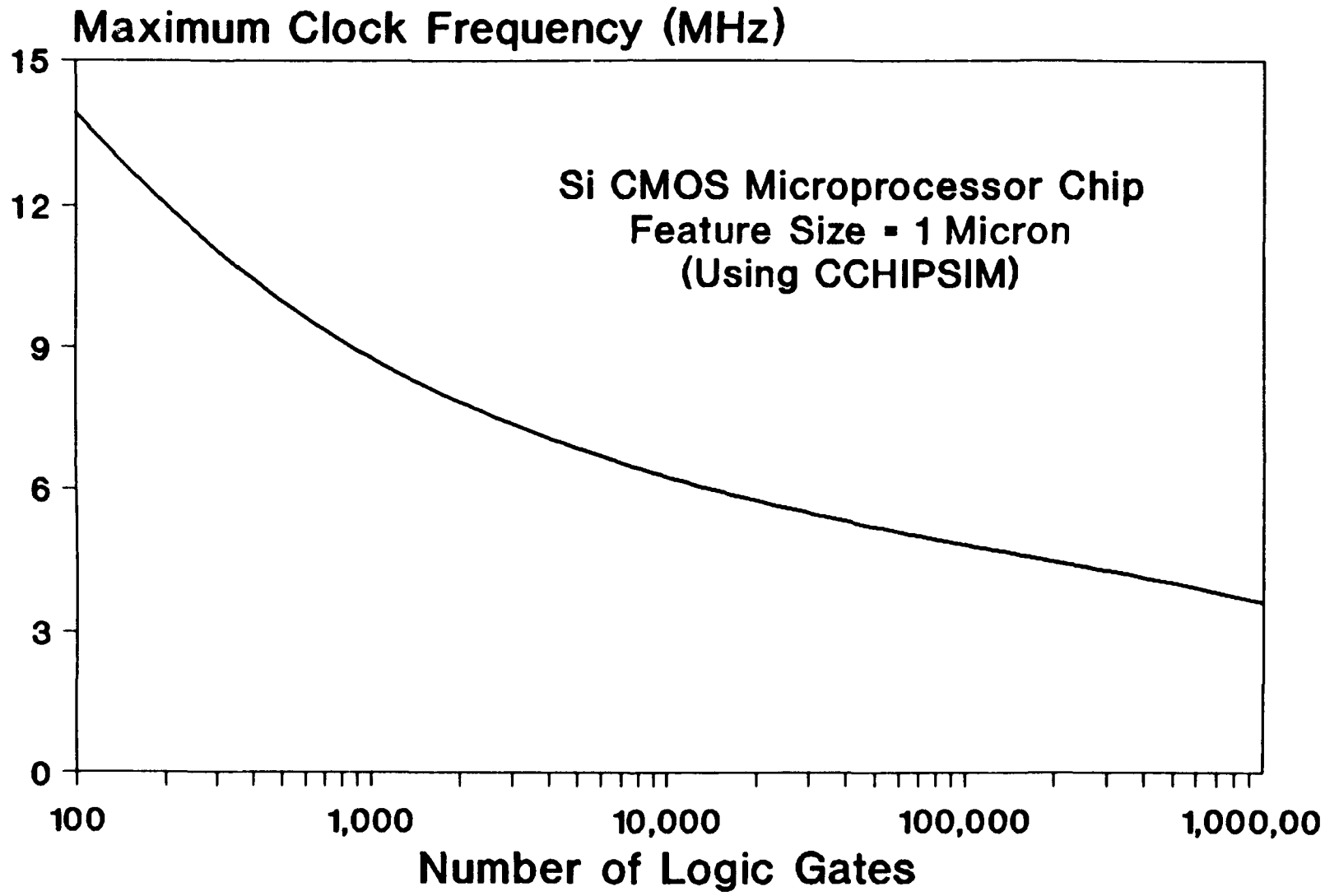

Maximum Clock Frequency (MHz)
15
12
9
6
3
0
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 21

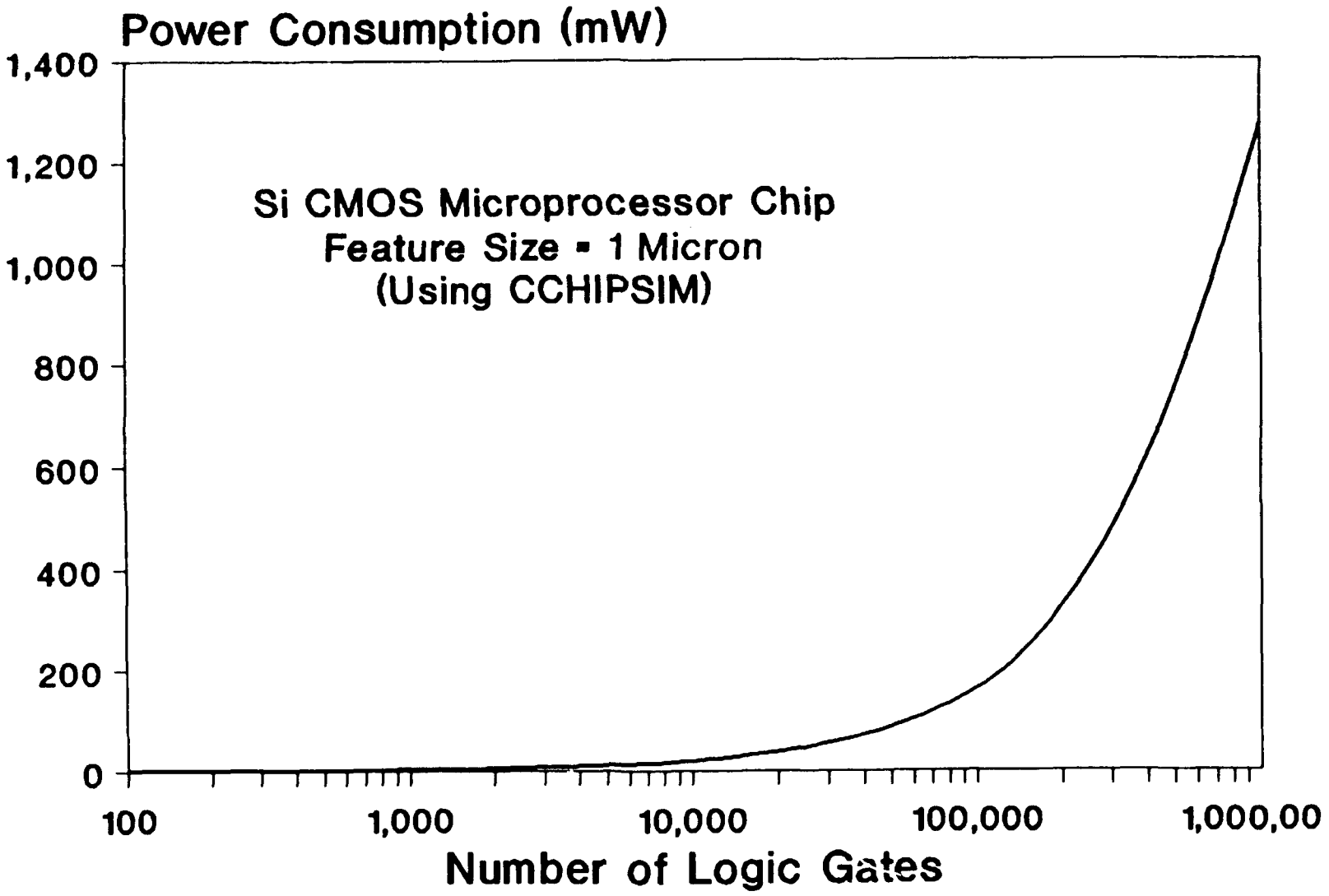

Power Consumption (mW)
1,400
1,200
1,000
800
600
400
200
0
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 22

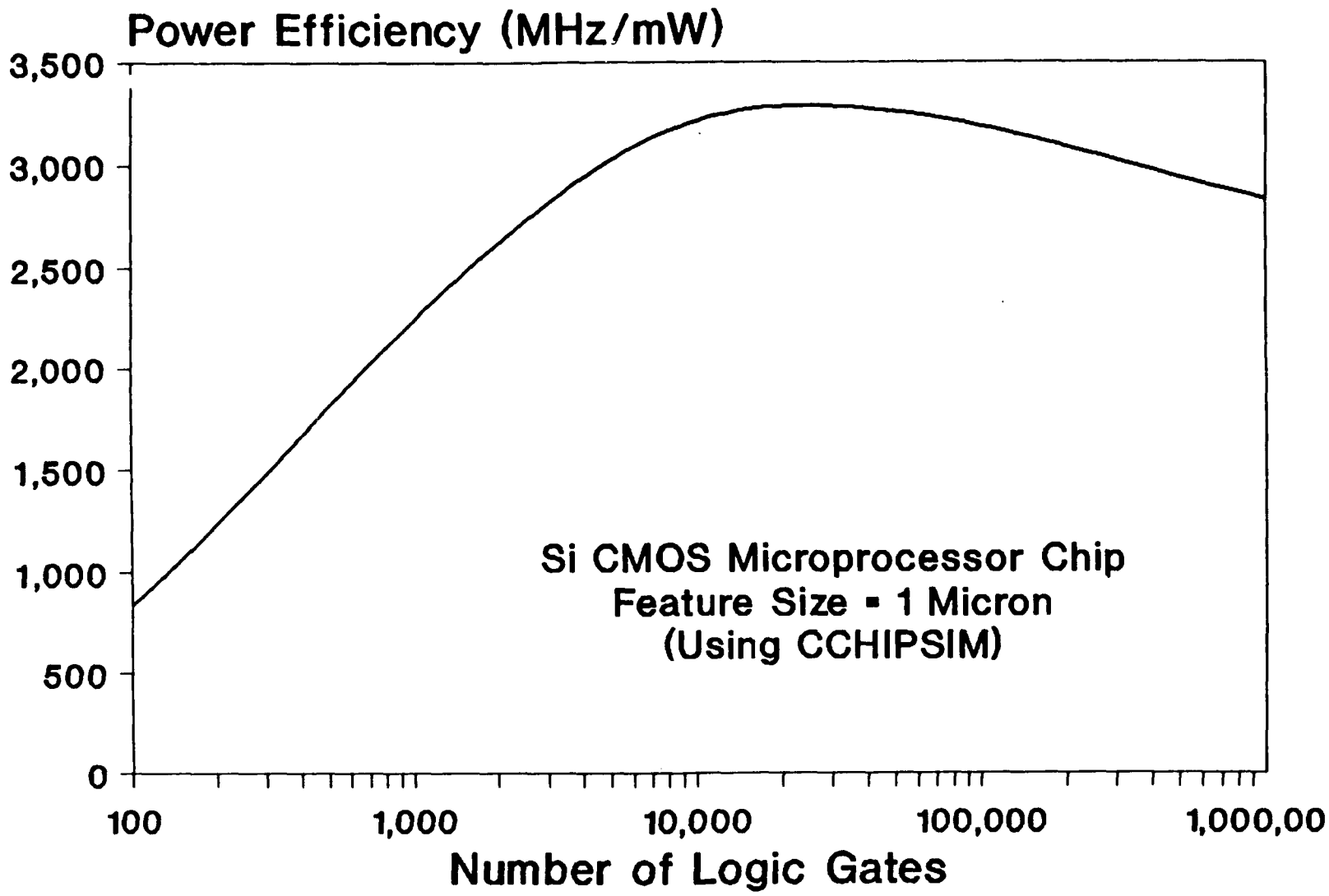

Power Efficiency (MHz/mW)
3,500
3,000
2,500
2,000
1,500
1,000
500
0
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)

FIGURE 23

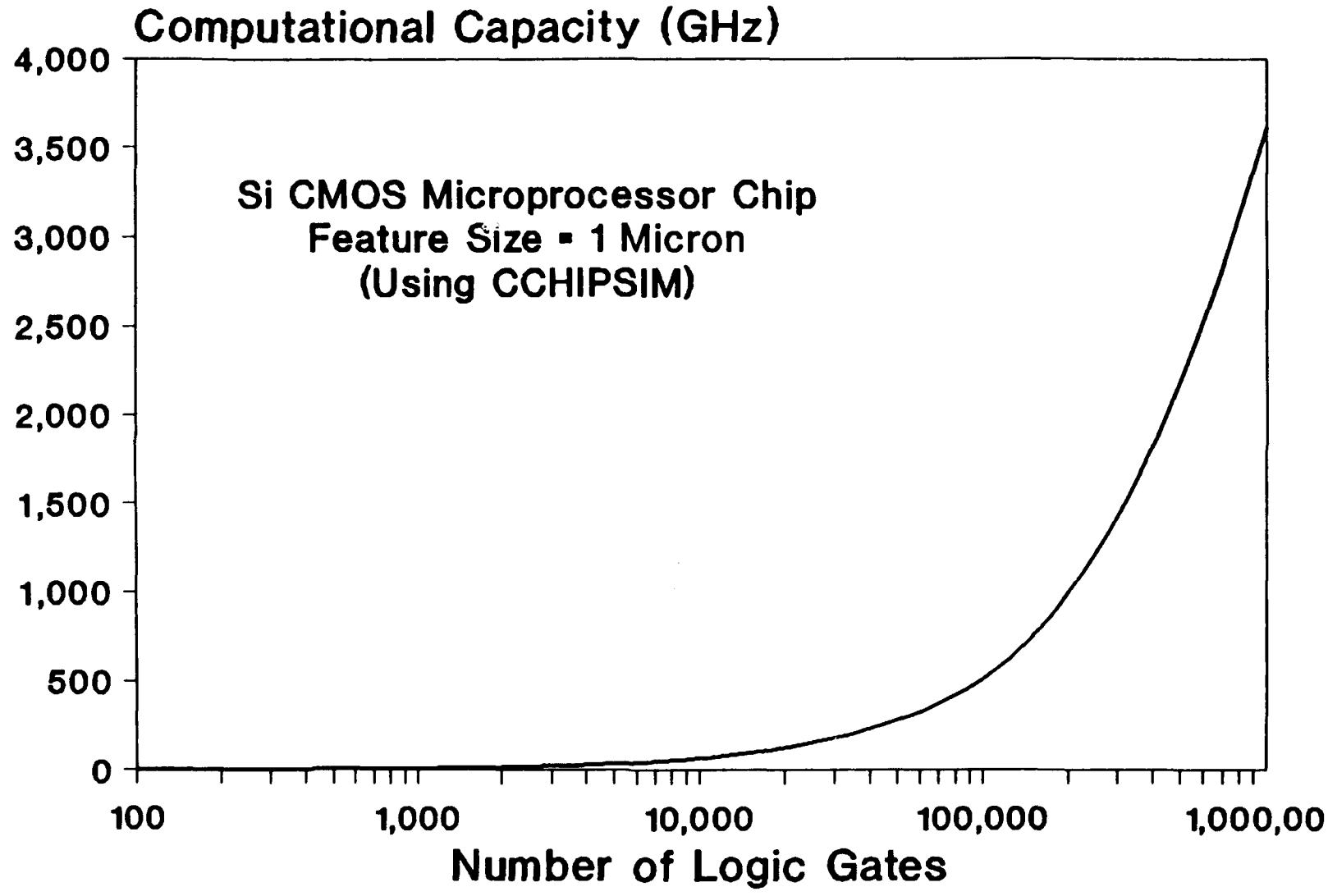

Computational Capacity (GHz)
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 24

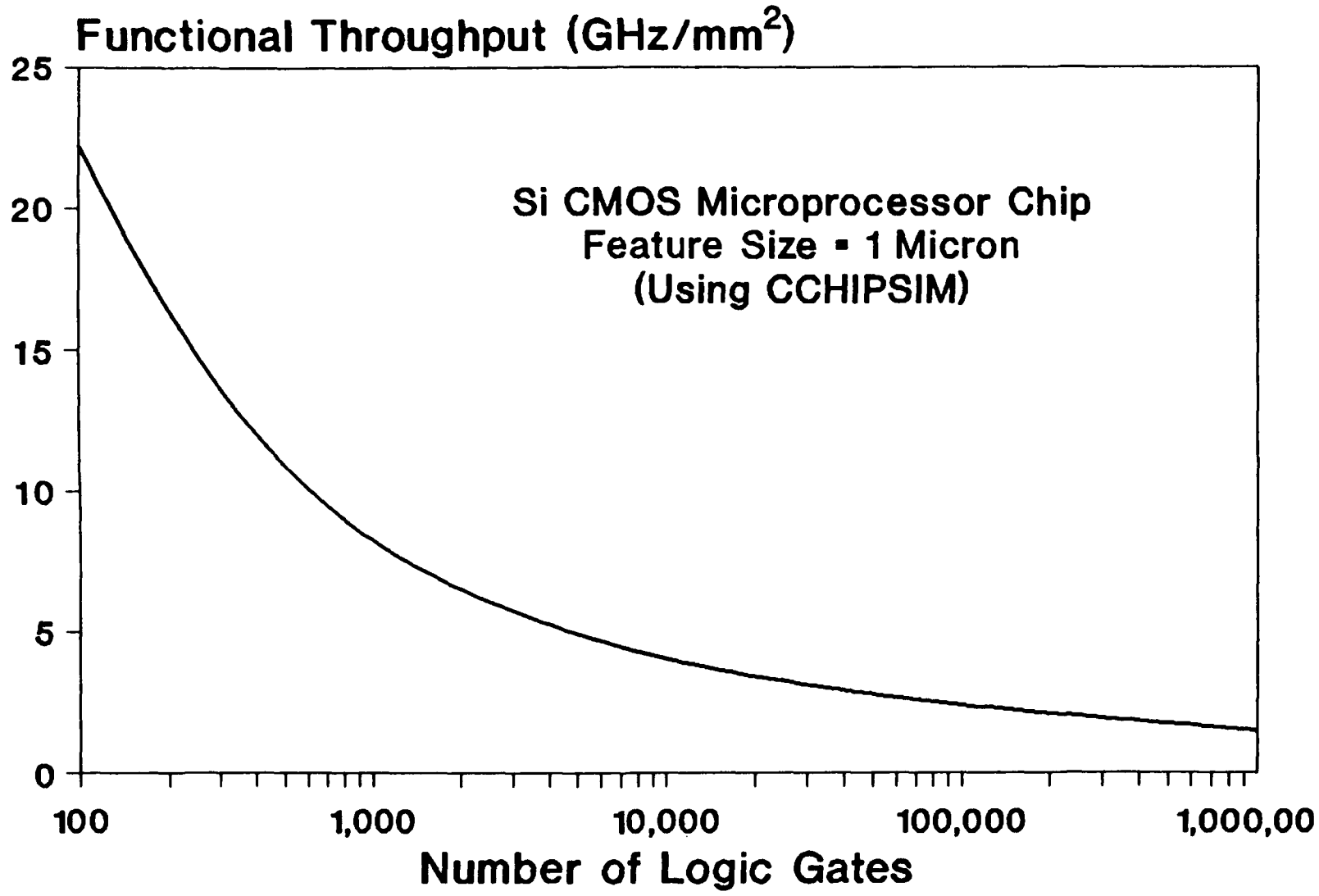

Functional Throughput (GHz/mm^2)
25
20
15
10
5
0
Si CMOS Microprocessor Chip
Feature Size = 1 Micron
(Using CCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 25

TABLE 5: Actual Data and Simulation Results for Si CMOS Microprocessor Chip Fairchild Clipper (1985)

Performance Indicator	Actual Value	Using CCHIPSIM	
Chip Size	10.0	12.3	(mm)
Maximum Clock Freq.	16.5	15.97	(MHz)
Power Consumption	0.5	0.8	(Watt)
Computational Cap.	*	351.4	(GHz)
Power Efficiency	*	0.4	(GHz/mW)
Functional Throughput	*	2.33	(GHz/mm^2)
Fabrication Yield	*	11.71	(%)

Chip Parameters	
Minimum Feature Size	= 2 microns
Number of Transistors	= 132,000
Interconnect Layers	= 3
Number of Pins	= 132
Interconnection Pitch	= 6 microns

* Data is not available

TABLE 6: Actual Data and Simulation Results for Si CMOS Microprocessor Chip Intel 80386 (1985)

Performance Indicator	Actual Value	Using CCHIPSIM
Chip Size	9.6	11.6 (mm)
Maximum Clock Freq.	12-16	15.41 (MHz)
Power Consumption	1-2	1.12 (Watt)
Computational Cap.	*	462.4 (GHz)
Power Efficiency	*	0.41 (GHz/mW)
Functional Throughput	*	2.17 (GHz/mm^2)
Fabrication Yield	*	8.6 (%)
Chip Parameters		
Minimum Feature Size	= 2 microns	
Number of Transistors	= 180,000	
Interconnect Layers	= 3	
Number of Pins	= 120	
Interconnection Pitch	= 6 microns	

* Data is not available

6. THE PROGRAM "BCHIPSIM" FOR SILICON BIPOLAR CHIPS

A microcomputer program called "BCHIPSIM" has been developed which can be used to predict the performance indicators of a microprocessor, gate array or a high speed computer chip based on the silicon bipolar technology as well as to study the dependence of these indicators on the feature size of the transistors and the integration level of the chip. The default values of the various chip, gate and the interconnection parameters used in BCHIPSIM are as follows:

Chip-Type Dependent Parameters for a Microprocessor Chip

Interconnect-length Rent's constant	=	0.4
Pin-count Rent's constant	=	0.45
Pin-count multiplication constant	=	0.82
Logic depth	=	22

Chip-Type Dependent Parameters for a Gate-Array Chip

Interconnect-length Rent's constant	=	0.5
Pin-count Rent's constant	=	0.5
Pin-count multiplication constant	=	1.9
Logic depth	=	30

Chip-Type Dependent Parameters for a High-Speed Computer Chip

Interconnect-length Rent's constant	=	0.6
Pin-count Rent's constant	=	0.63

Pin-Count multiplication constant	=	1.4
Logic depth	=	10

Chip Parameters

Number of logic gates on the chip	=	10,000
Fan-out of a typical gate on the chip	=	3
Total current at an output buffer	=	1 mA
Fraction of on-chip gates that switch during a clock cycle	=	0.3
Density of defects on the chip	=	$5/cm^2$
Number of pins on the chip	=	190

Gate Parameters

Minimum feature size	=	1 μm
Base-emitter capacitance of feature size transistor	=	20 fF
Logic swing	=	0.5 volt
Gate current source	=	0.25 mA
Gate power supply voltage	=	4 volts
Ratio of optimum-size to feature-size transistors	=	1
Collector-base capacitance	=	5 fF
Collector-substrate capacitance	=	6 fF
Transistor base resistance	=	500 Ω
Transistor base delay	=	3 ps

Time delay in a typical gate	=	264 ps

Interconnection Parameters

Number of interconnection layers	=	4
Widths of on-chip interconnects	=	2 μm
Pitches of on-chip interconnects	=	4 μm
Thicknesses of on-chip interconnects	=	0.5 μm
Thickness of the dielectric material	=	0.5 μm
Utilization coefficient of interconnections	=	0.4
Interconnection resistance	=	28 Ω/mm
Interconnection capacitance	=	0.27 pF/mm

6.1 SIMULATION RESULTS USING BCHIPSIM

BCHIPSIM has been used to predict the dependence of the chip performance on its minimum feature size as well as on its integration level. For example, the dependences of the chip size, maximum clock frequency, power consumption, computational capacity, power efficiency and functional throughput rate of a 10,000-logic gates silicon bipolar single-chip gate-array on its minimum feature size in the range 0.1 to 5 μm are shown in figures 26 to 31, respectively; and the dependences of each of these performance indicators for a 1 μm silicon bipolar gate-array chip on its integration level in the range 100 to 1,000,000 logic gates are shown in figures 32 to 37, respectively.

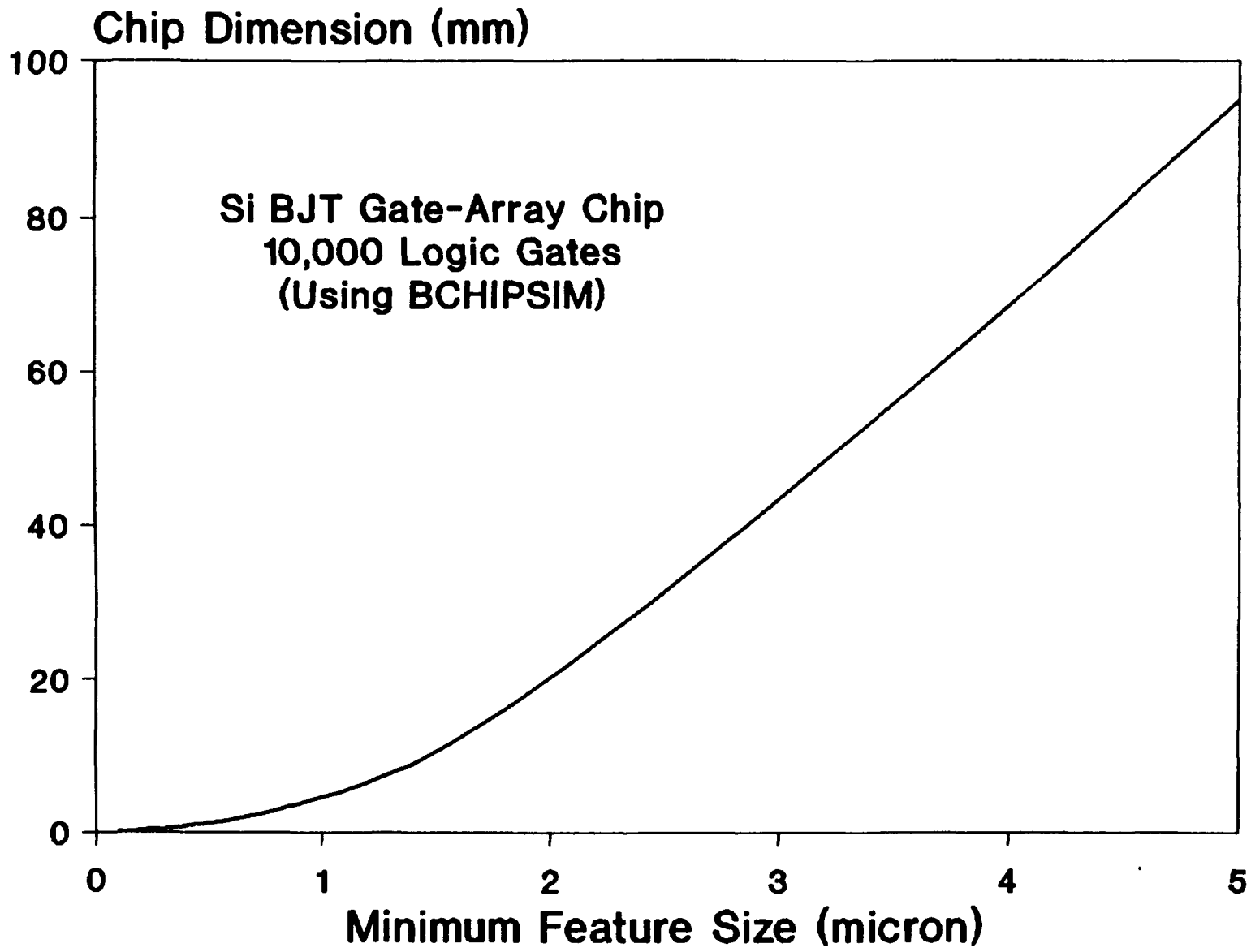

Chip Dimension (mm)
100
80
60
40
20
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 26

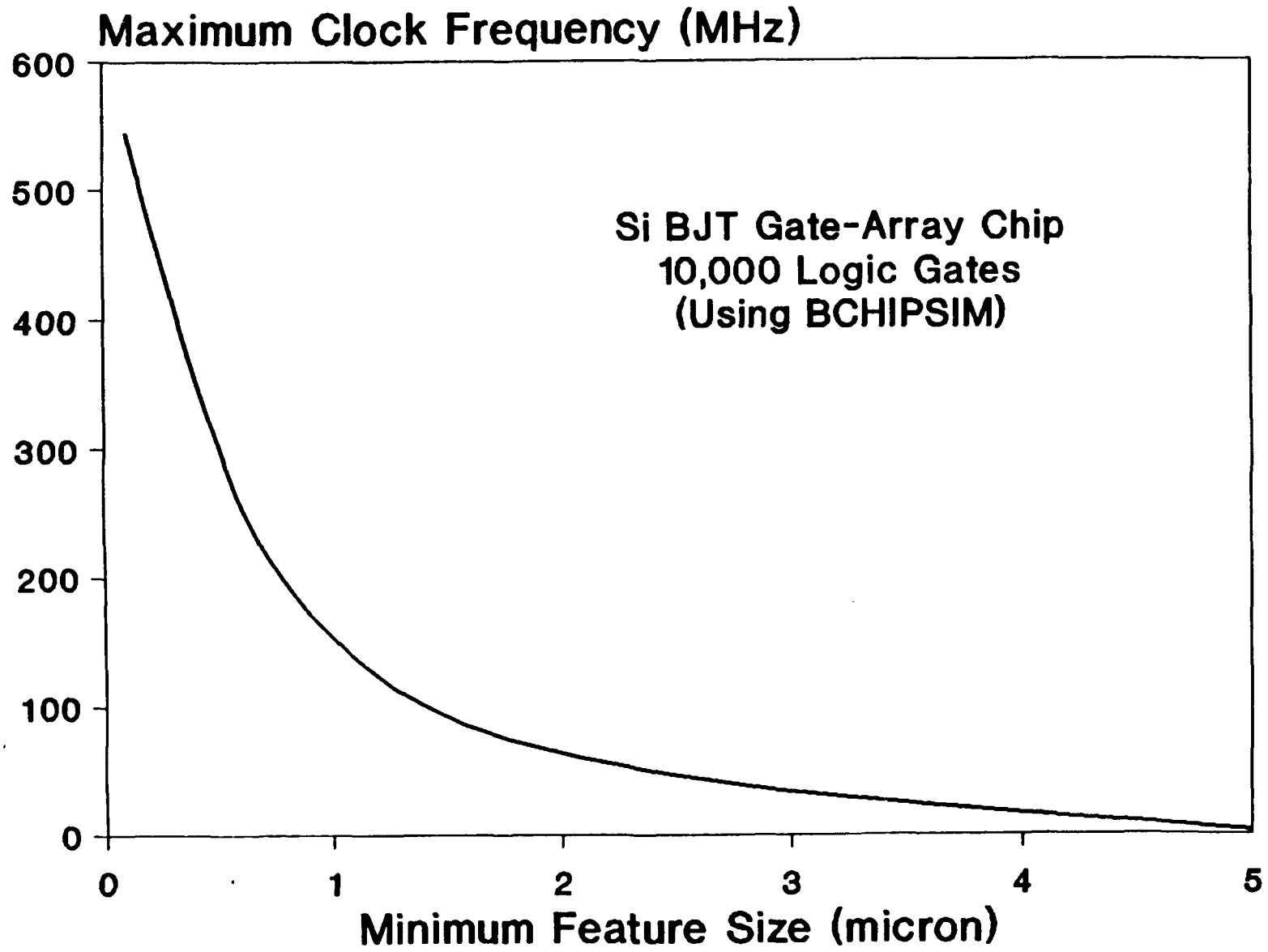

Maximum Clock Frequency (MHz)
600
500
400
300
200
100
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 27

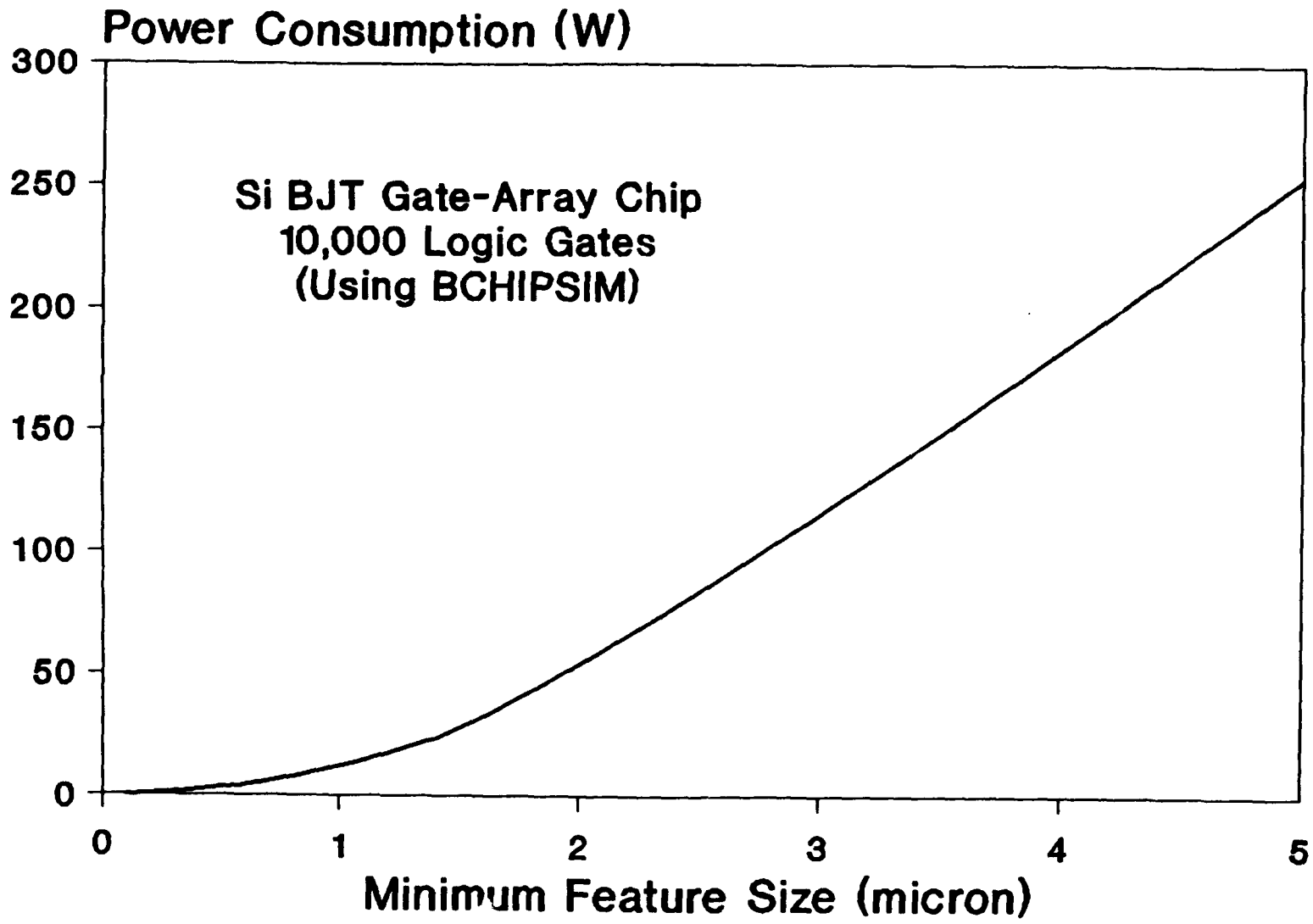

Power Consumption (W)
300
250
200
150
100
50
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 28

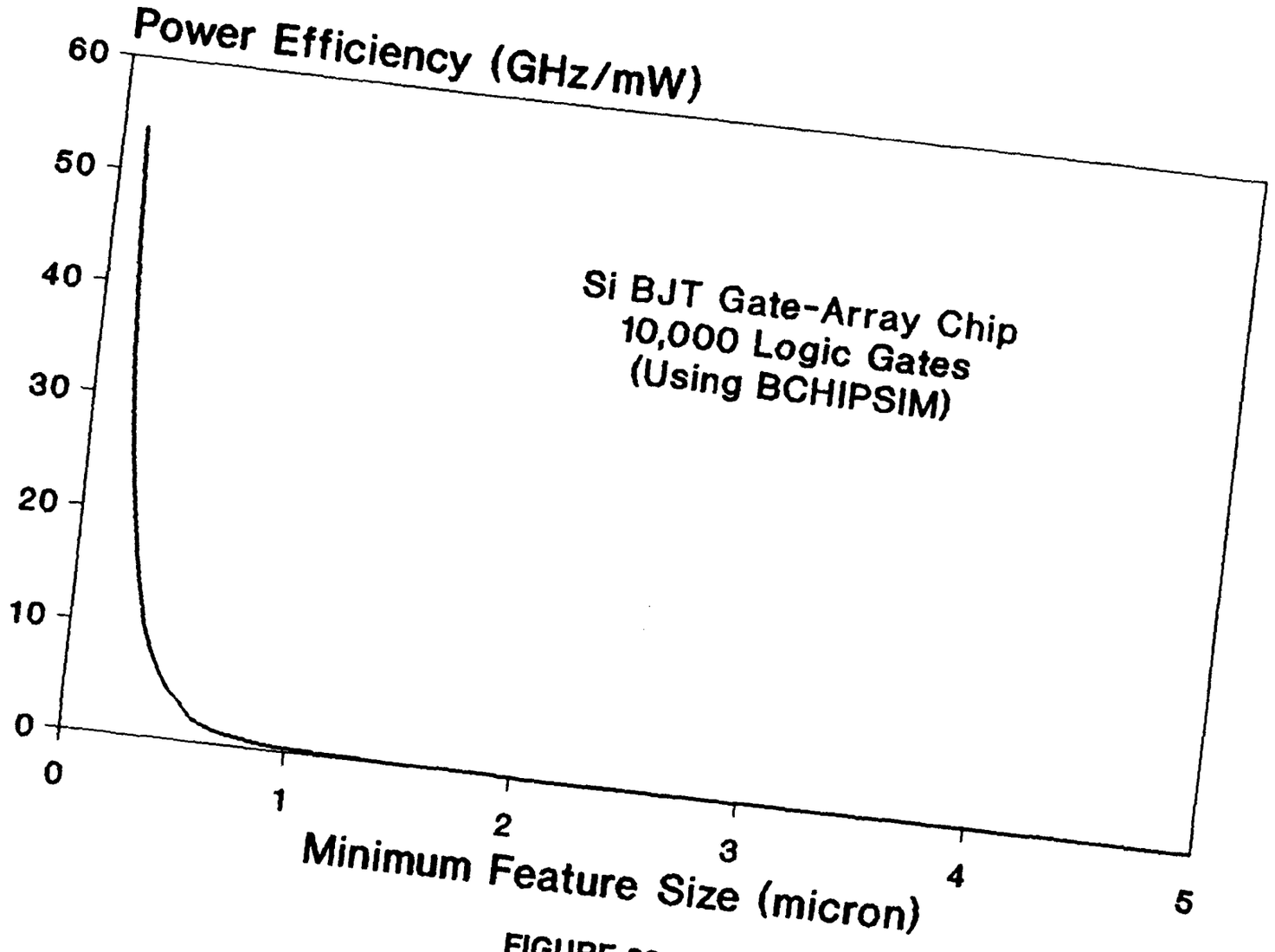

Power Efficiency (GHz/mW)
60
50
40
30
20
10
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 29

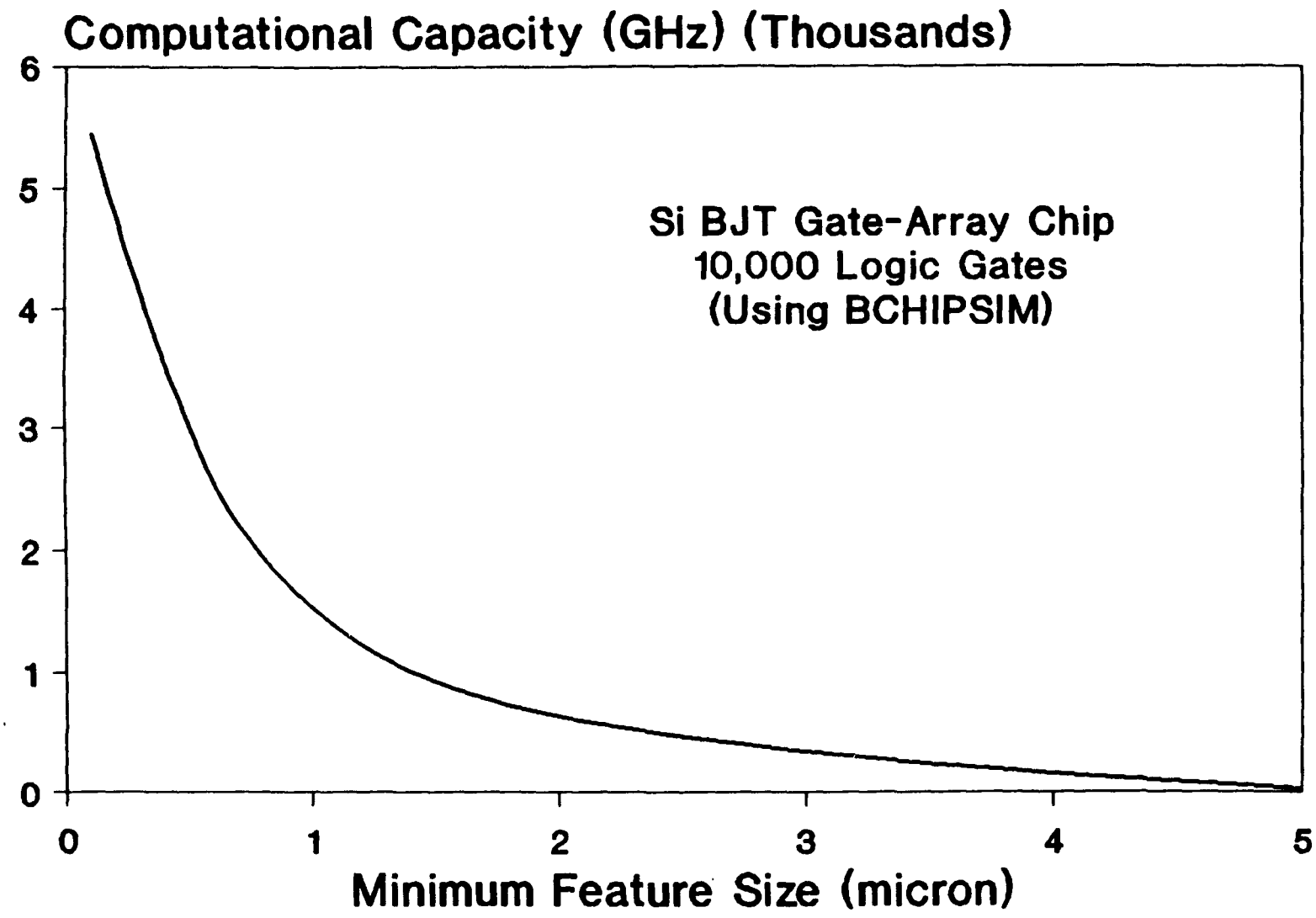

Computational Capacity (GHz) (Thousands)
6
5
4
3
2
1
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 30

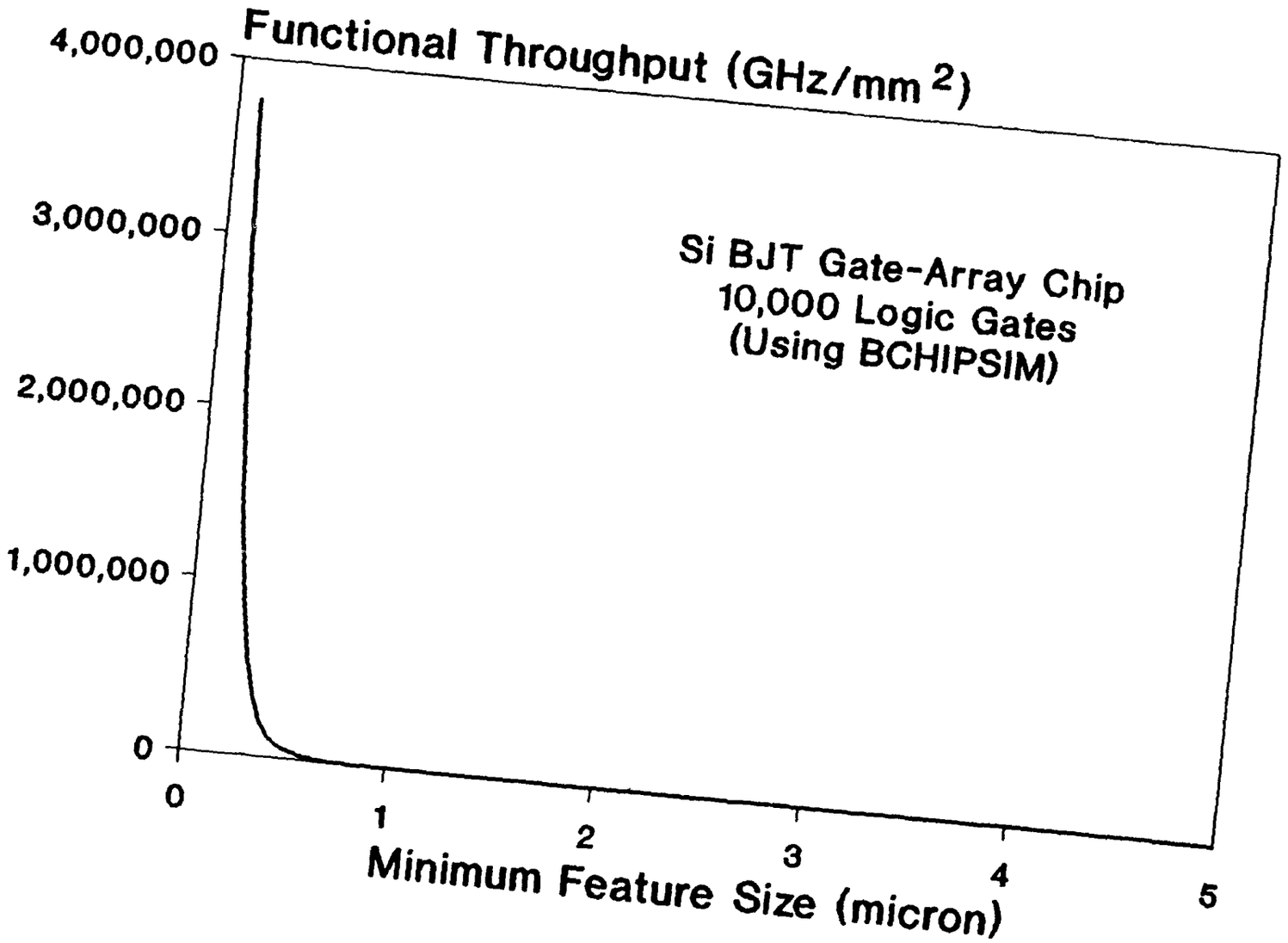

Functional Throughput (GHz/mm 2)
4,000,000
3,000,000
2,000,000
1,000,000
0
Si BJT Gate-Array Chip
10,000 Logic Gates
(Using BCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 31

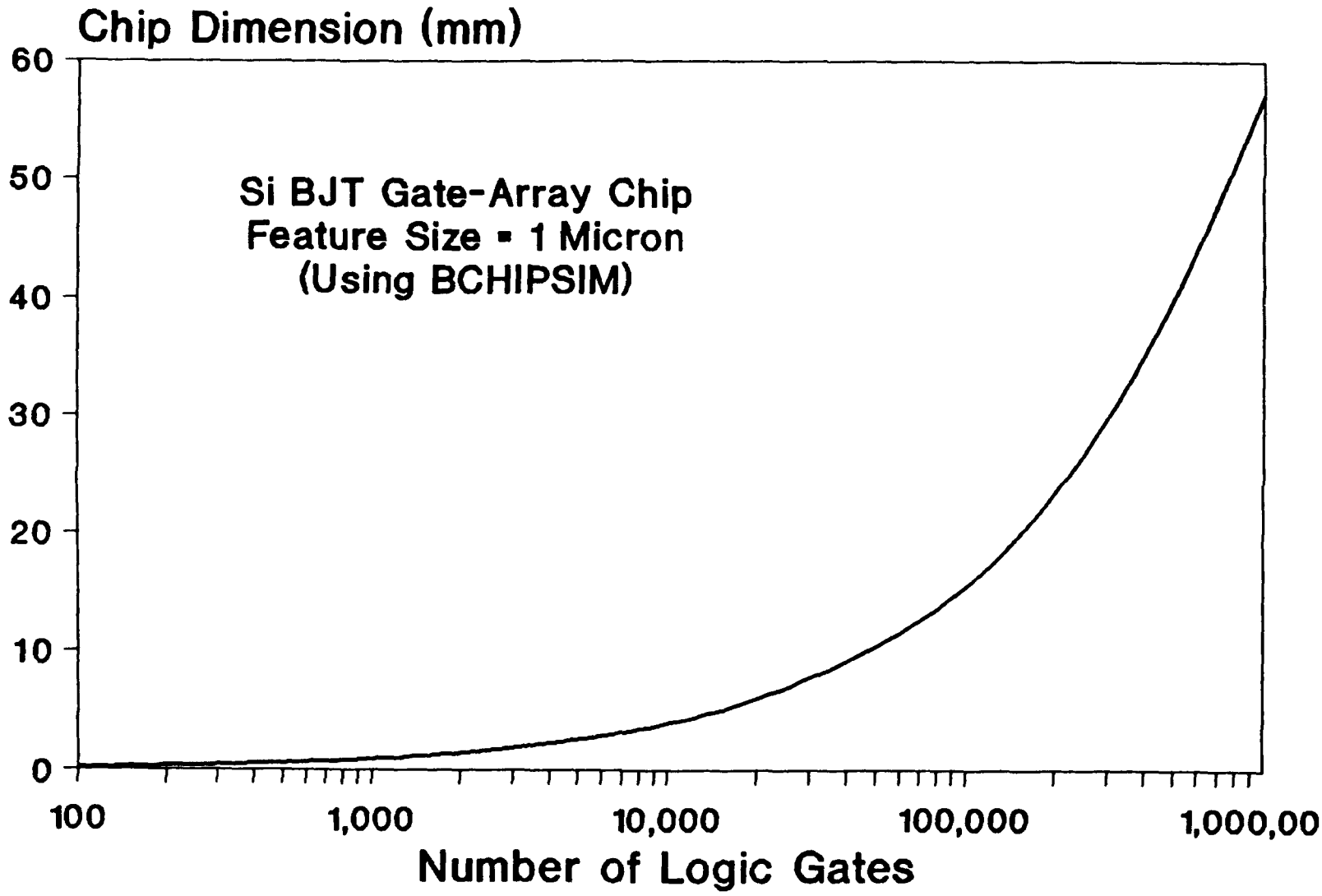

Chip Dimension (mm)
60
50
40
30
20
10
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 32

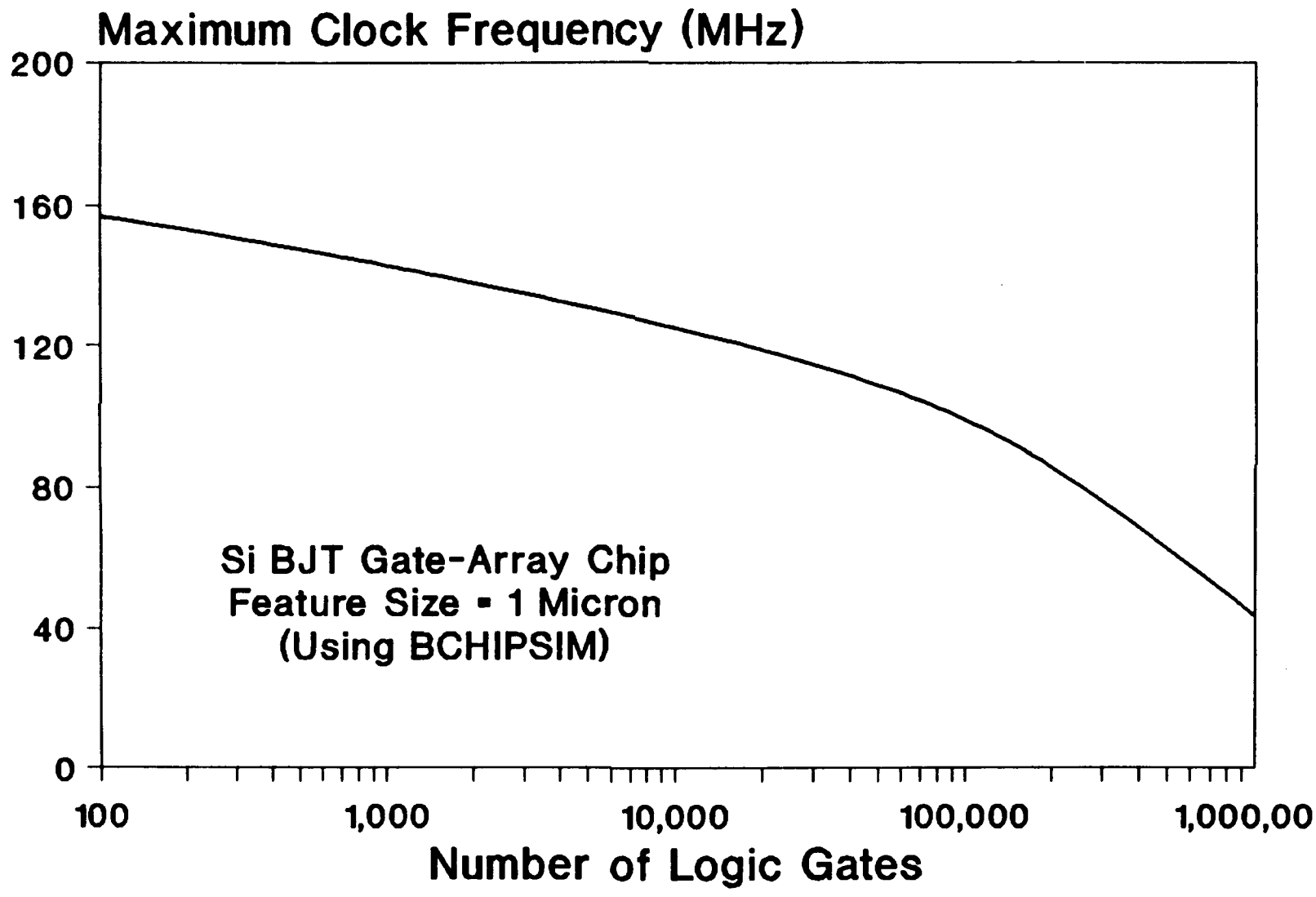

Maximum Clock Frequency (MHz)
200
160
120
80
40
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 33

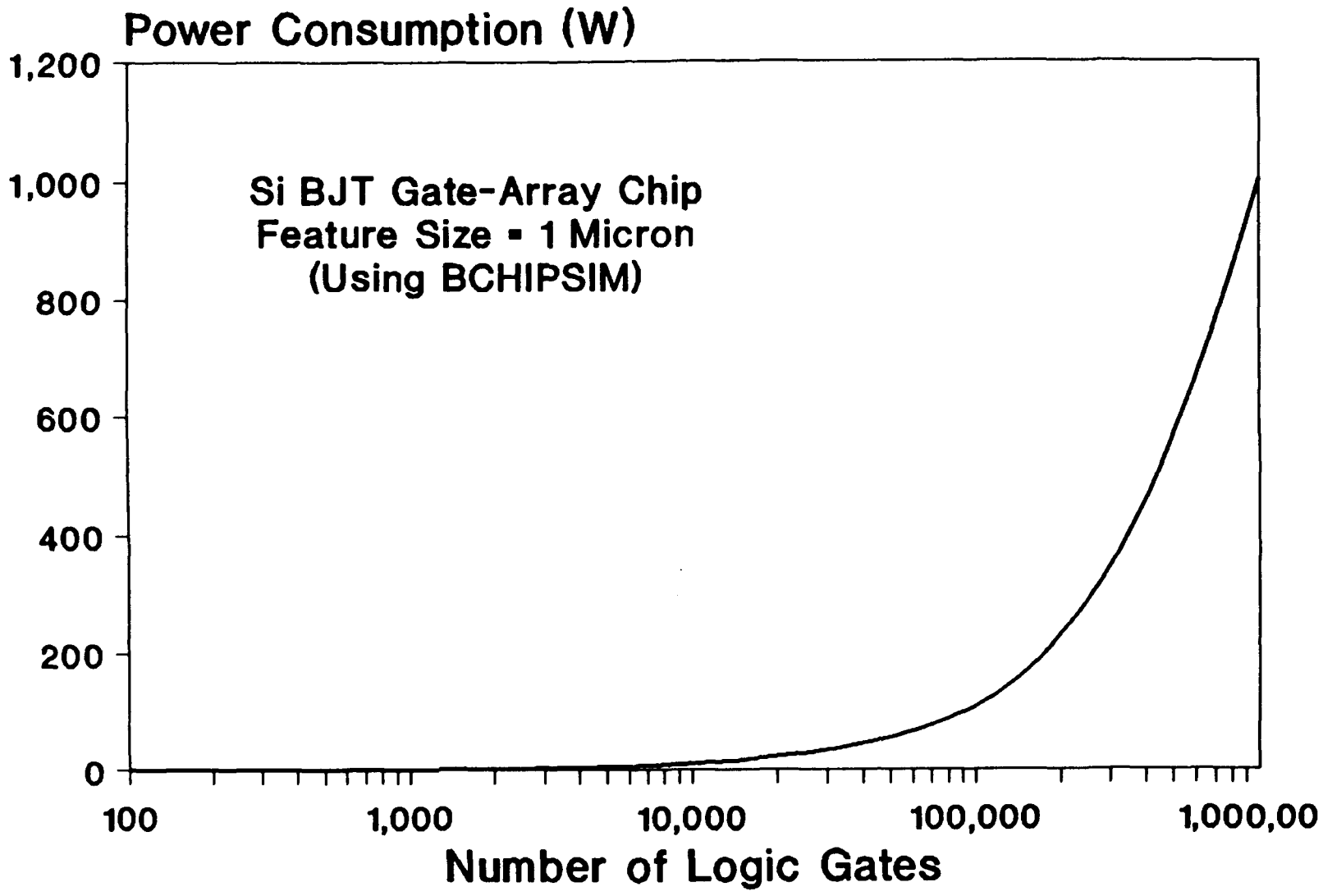

Power Consumption (W)
1,200
1,000
800
600
400
200
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 34

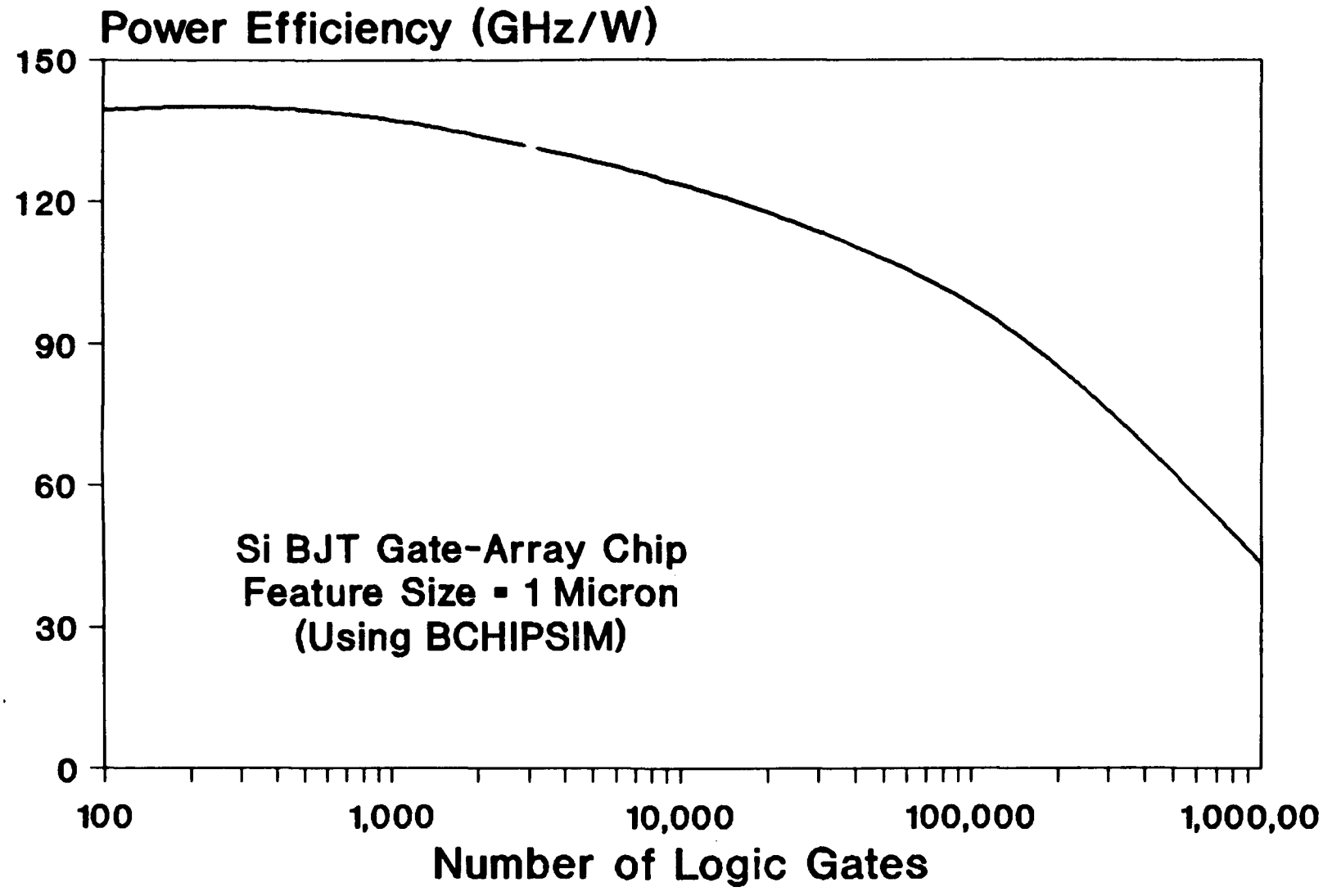

Power Efficiency (GHz/W)
150
120
90
60
30
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 35

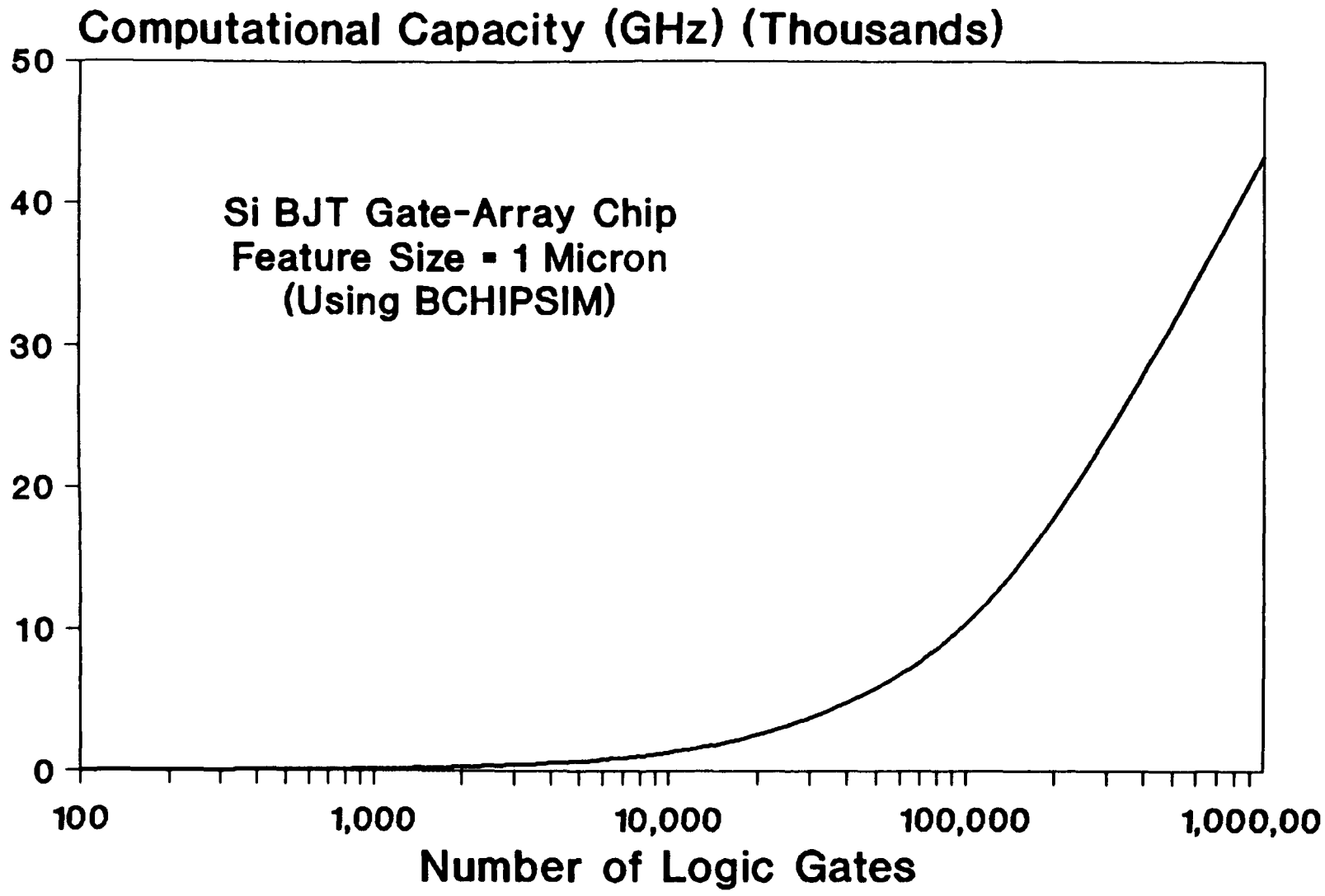

Computational Capacity (GHz) (Thousands)
50
40
30
20
10
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 36

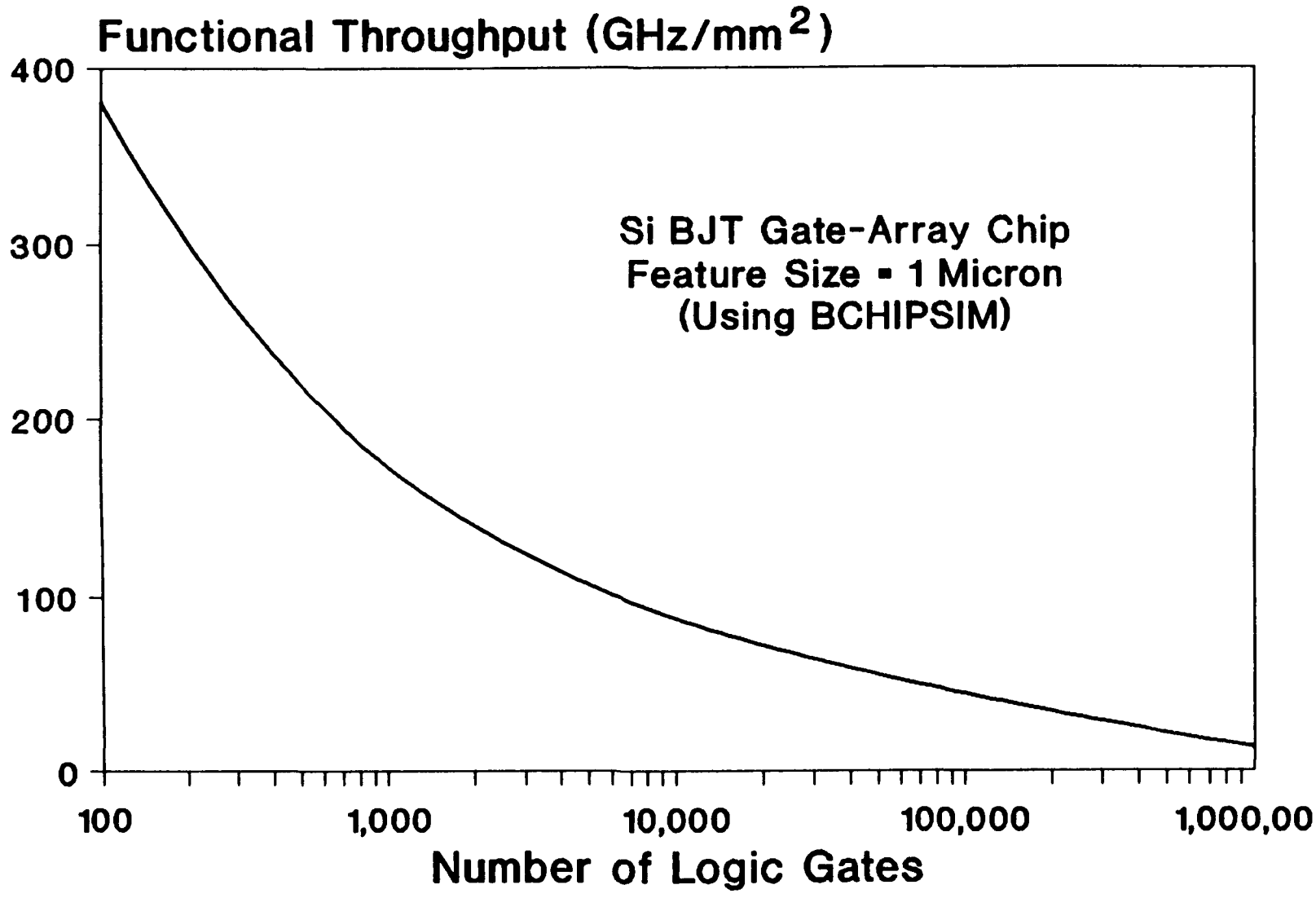

Functional Throughput (GHz/mm^2)
400
300
200
100
0
Si BJT Gate-Array Chip
Feature Size = 1 Micron
(Using BCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 37

7. THE PROGRAM "GCHIPSIM" FOR GaAs HETEROJUNCTION BIPOLAR CHIPS

A microcomputer program called "GCHIPSIM" has been developed which can be used to predict the performance indicators of a conventional microprocessor, a conventional gate array or a high-speed computer chip based on the GaAs HBT technology as well as to study the dependence of these indicators on the feature size of the transistors and the integration level of the chip. In this program, the interconnection capacitances were calculated by the method of moments in conjunction with a Green's function appropriate for the geometry of the interconnections [8]. The capacitance results included the fringing fields as well as the shielding effects of neighboring interconnections for high-density chips. The default values of the various chip, gate and the interconnection parameters used in GCHIPSIM are given below. Some of these were derived from the values provided by Dr. Ross Mactaggard of Honeywell, Inc. [9].

Chip-Type Dependent Parameters for a Microprocessor Chip

Interconnect-length Rent's constant	=	0.4
Pin-count Rent's constant	=	0.45
Pin-count multiplication constant	=	0.82
Logic depth	=	22

Chip-Type Dependent Parameters for a Gate-Array Chip

Interconnect-length Rent's constant	=	0.5
Pin-count Rent's constant	=	0.5
Pin-count multiplication constant	=	1.9
Logic depth	=	30

Chip-Type Dependent Parameters for a High-Speed Computer Chip

Interconnect-length Rent's constant	=	0.6
Pin-count Rent's constant	=	0.63
Pin-count multiplication constant	=	1.4
Logic depth	=	10

Chip Parameters

Number of logic gates on the chip	=	10 ,000
Fan-out of a typical gate on the chip	=	3
Total current at an output buffer	=	1 mA
Fraction of on-chip gates that switch during a clock cycle	=	0.3
Density of defects on the chip	=	$5/cm^2$

Gate Parameters

Minimum feature size	=	1 μm
Total base-emitter capacitance of feature size transistor	=	5 fF
Base resistance of feature-size transistor	=	300 Ω
Gate power supply	=	4 volts
Ratio of optimum-size to feature-size transistors	=	1
Gate current source	=	0.25 mA
Logic swing	=	0.5 volt
Transistor base delay	=	0.5 ps

Collector-base capacitance	=	4 fF
Collector-substrate capacitance	=	0
Time delay in a typical gate	=	224 ps

Interconnection Parameters

Number of interconnection layers	=	4
Widths of on-chip interconnects	=	2 μm
Pitches of on-chip interconnects	=	4 μm
Thicknesses of on-chip interconnects	=	0.5 μm
Thickness of the GaAs substrate	=	350 μm
Utilization coefficient of interconnections	=	0.33
Interconnection resistance	=	28 Ω/mm
Interconnection capacitance	=	0.32 pF/mm

7.1 SIMULATION RESULTS USING GCHIPSIM

GCHIPSIM has been used to predict the dependence of the chip performance on its minimum feature size as well as on its integration level. For example, the dependences of the chip size, maximum clock frequency, power consumption, computational capacity, power efficiency and functional throughput rate of a 10,000-gate GaAs HBT high speed computer chip on its minimum feature size in the range 0.1 to 5.0 μm are shown in figures 38 to 43, respectively; and the dependences of each of these performance indicators for a 1-μm GaAs HBT high-speed computer chip on its integration level in the range 100 to 1,000,000 logic gates are shown in figures 44 to 49, respectively.

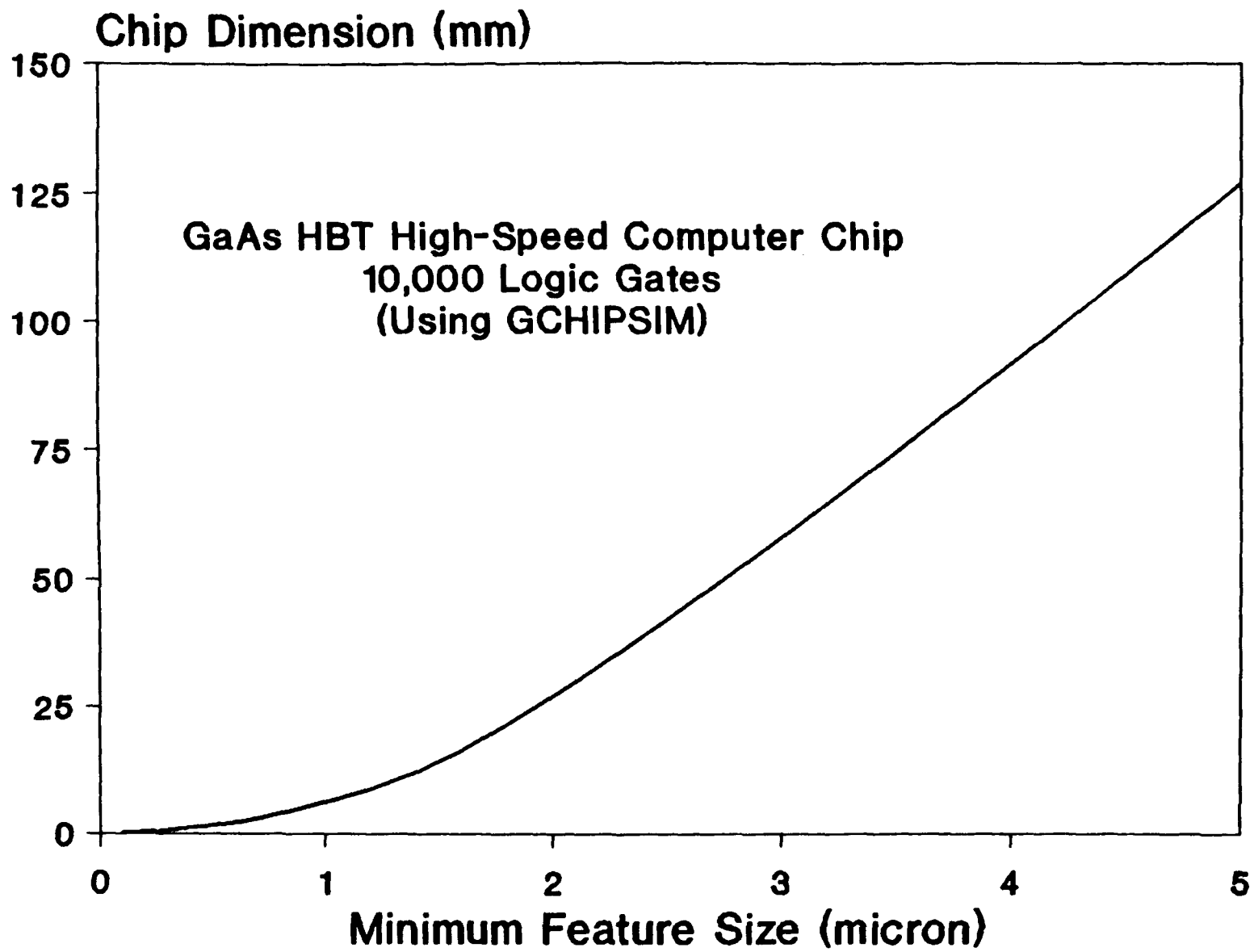

Chip Dimension (mm)
150
125
100
75
50
25
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 38

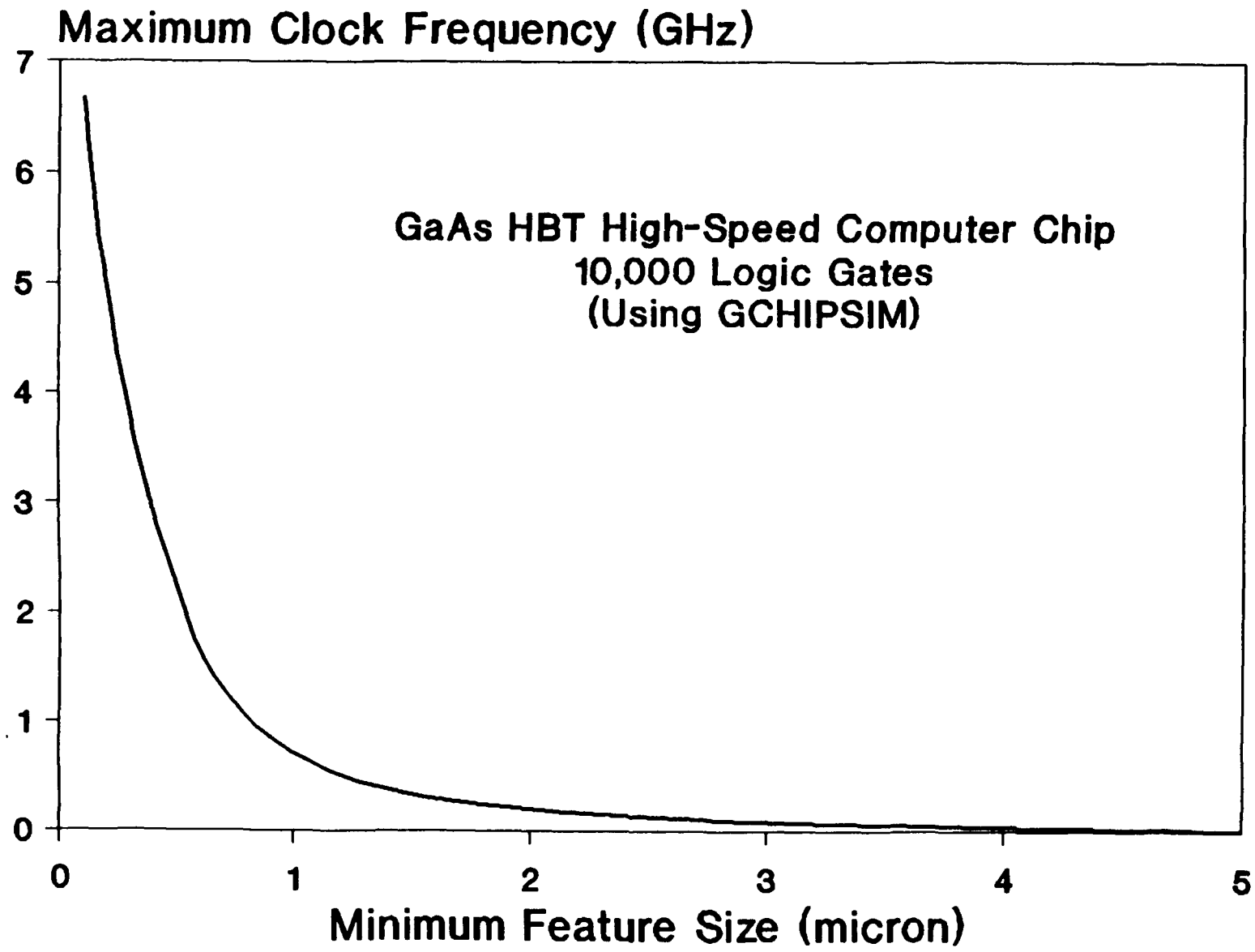

Maximum Clock Frequency (GHz)
7
6
5
4
3
2
1
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 39

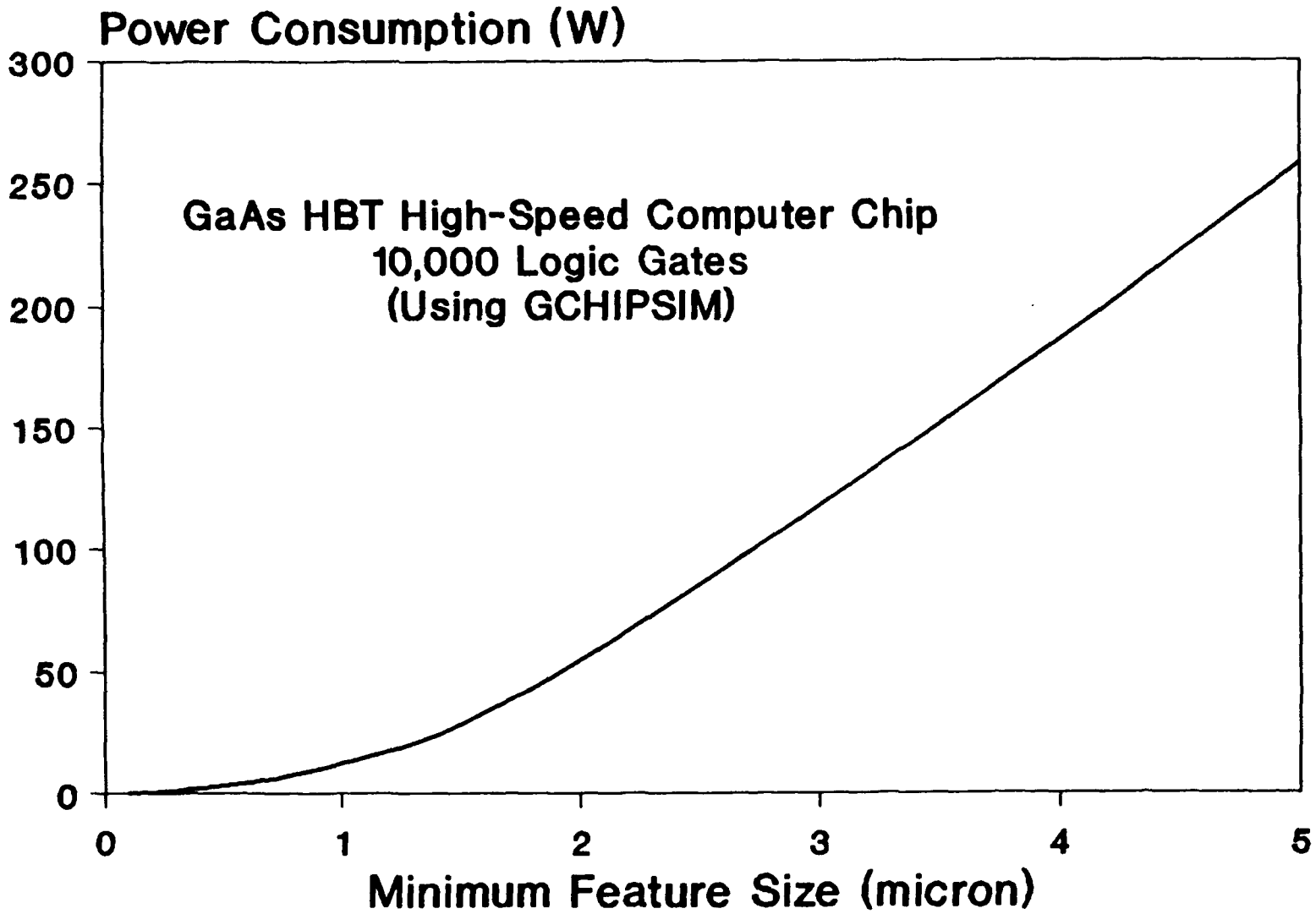

Power Consumption (W)
300
250
200
150
100
50
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 40

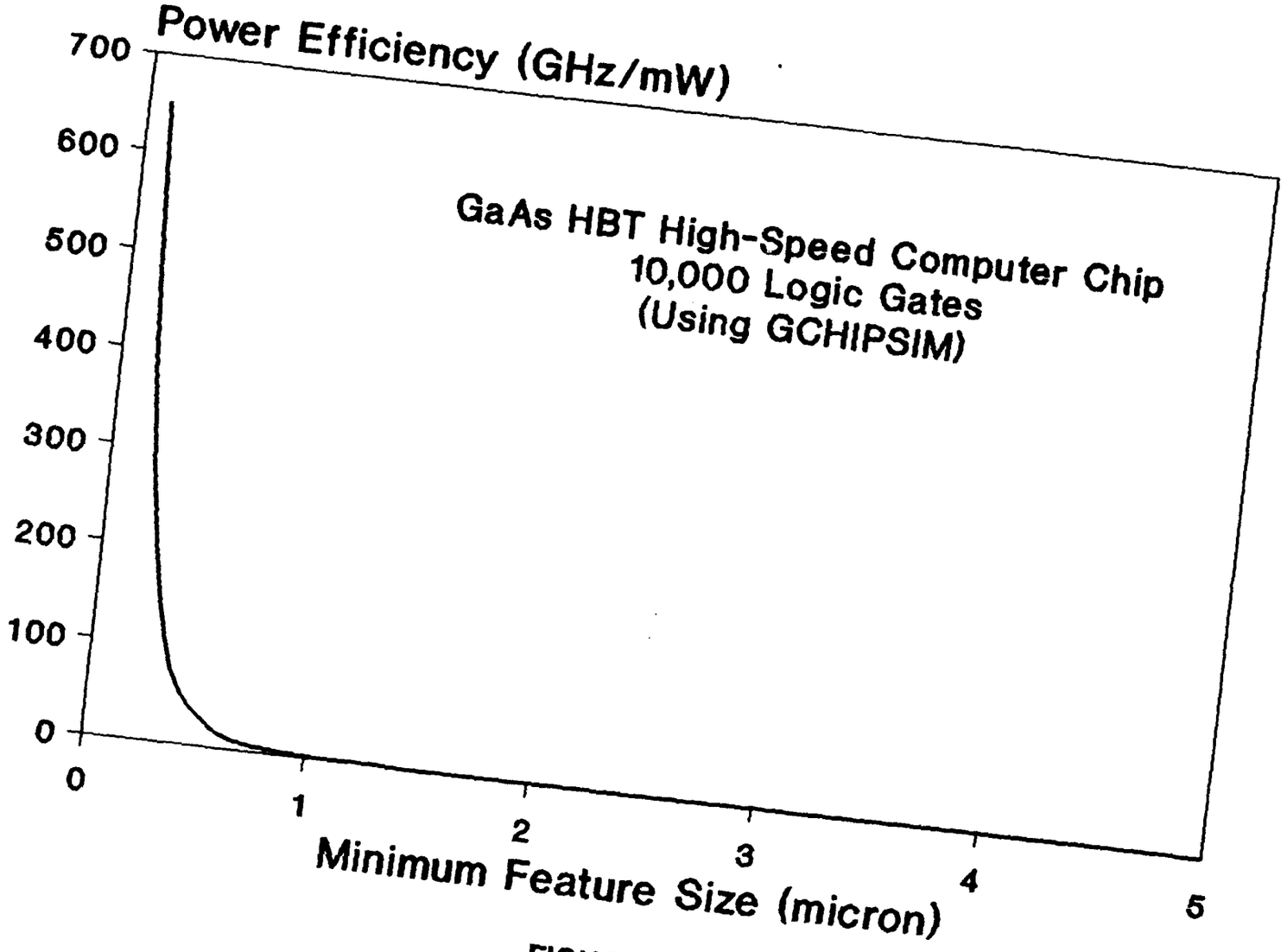

Power Efficiency (GHz/mW)
700
600
500
400
300
200
100
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 41

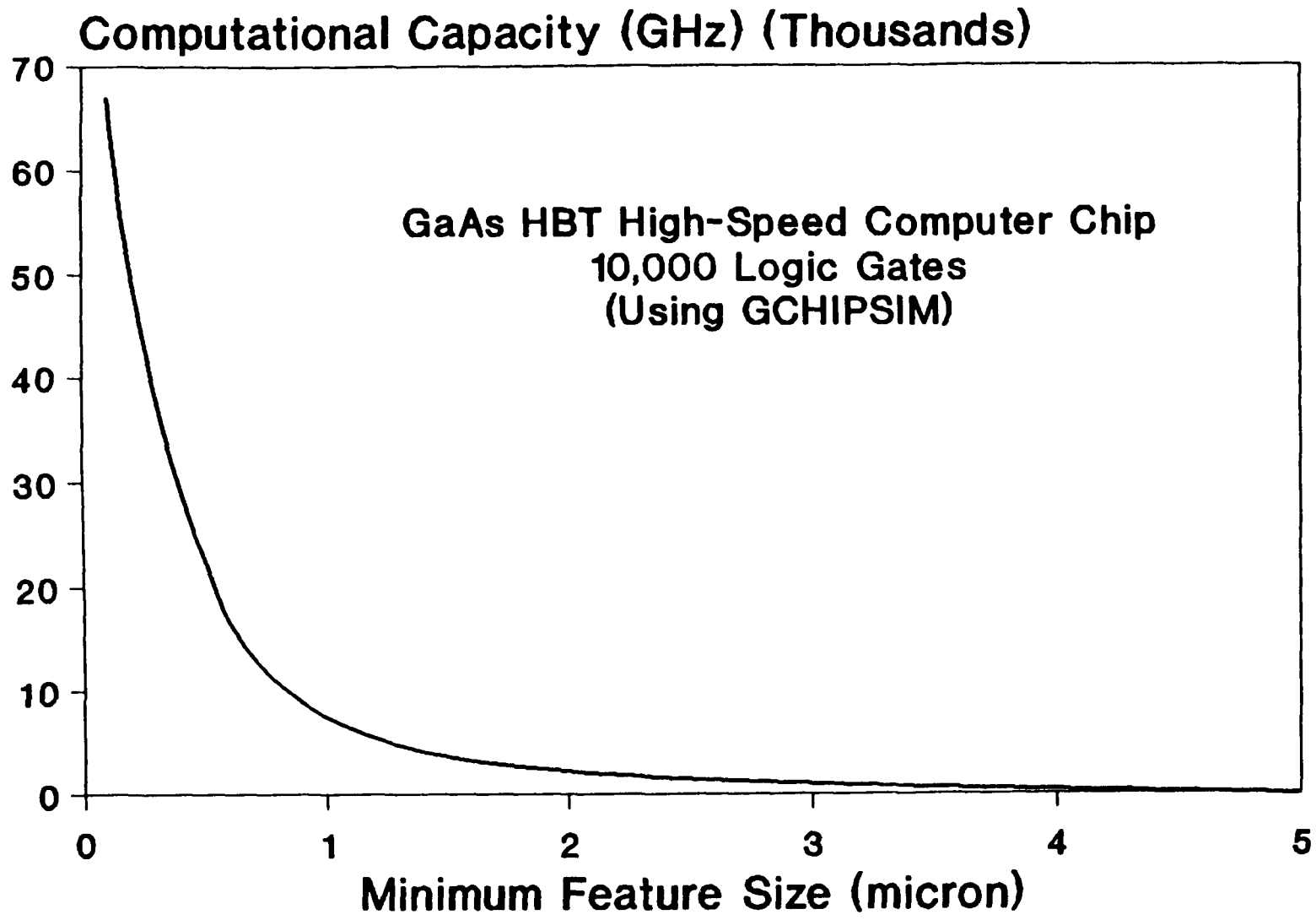

Computational Capacity (GHz) (Thousands)
70
60
50
40
30
20
10
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 42

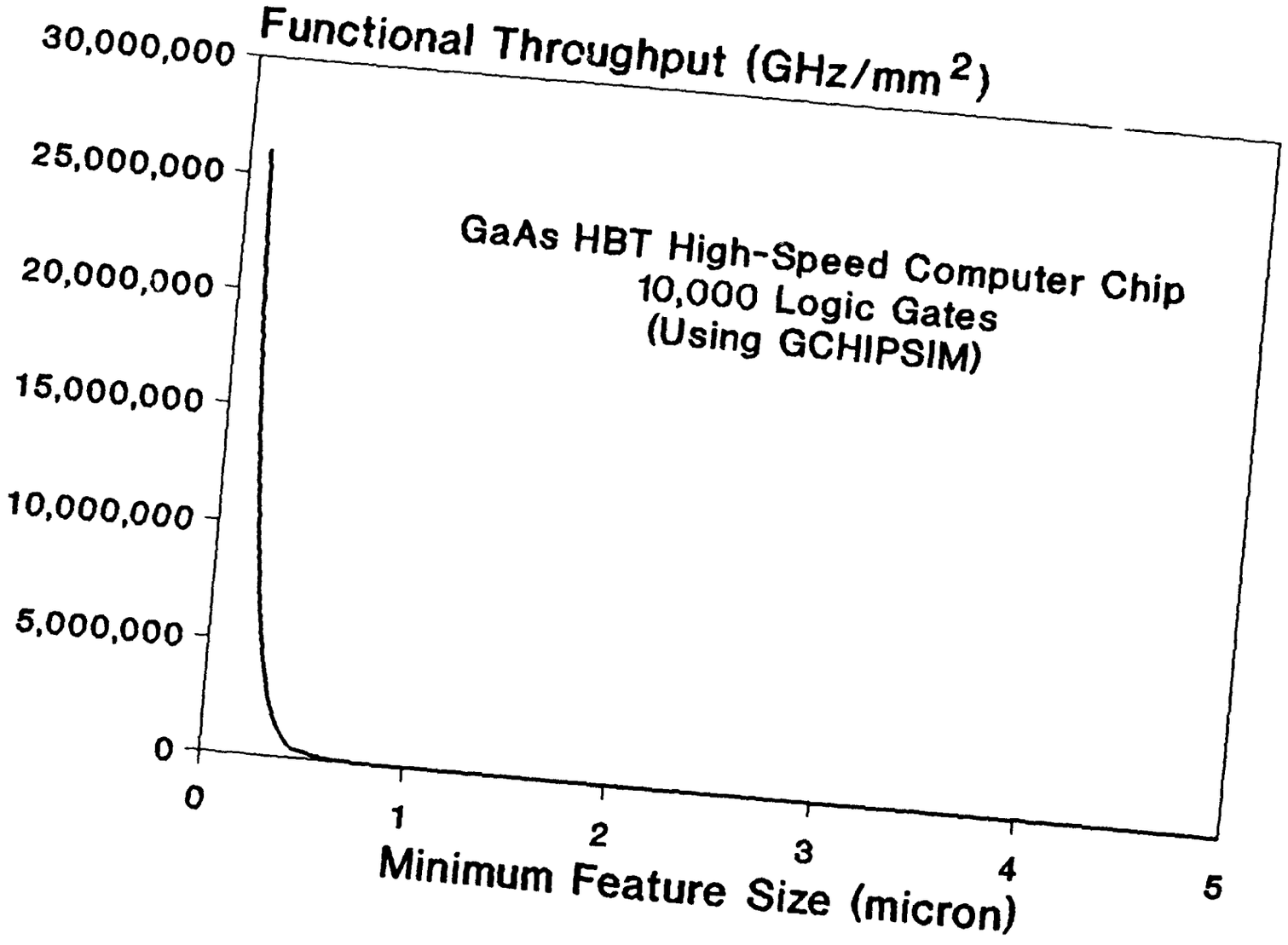

Functional Throughput (GHz/mm 2)
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
GaAs HBT High-Speed Computer Chip
10,000 Logic Gates
(Using GCHIPSIM)
0
1
2
3
4
5
Minimum Feature Size (micron)

FIGURE 43

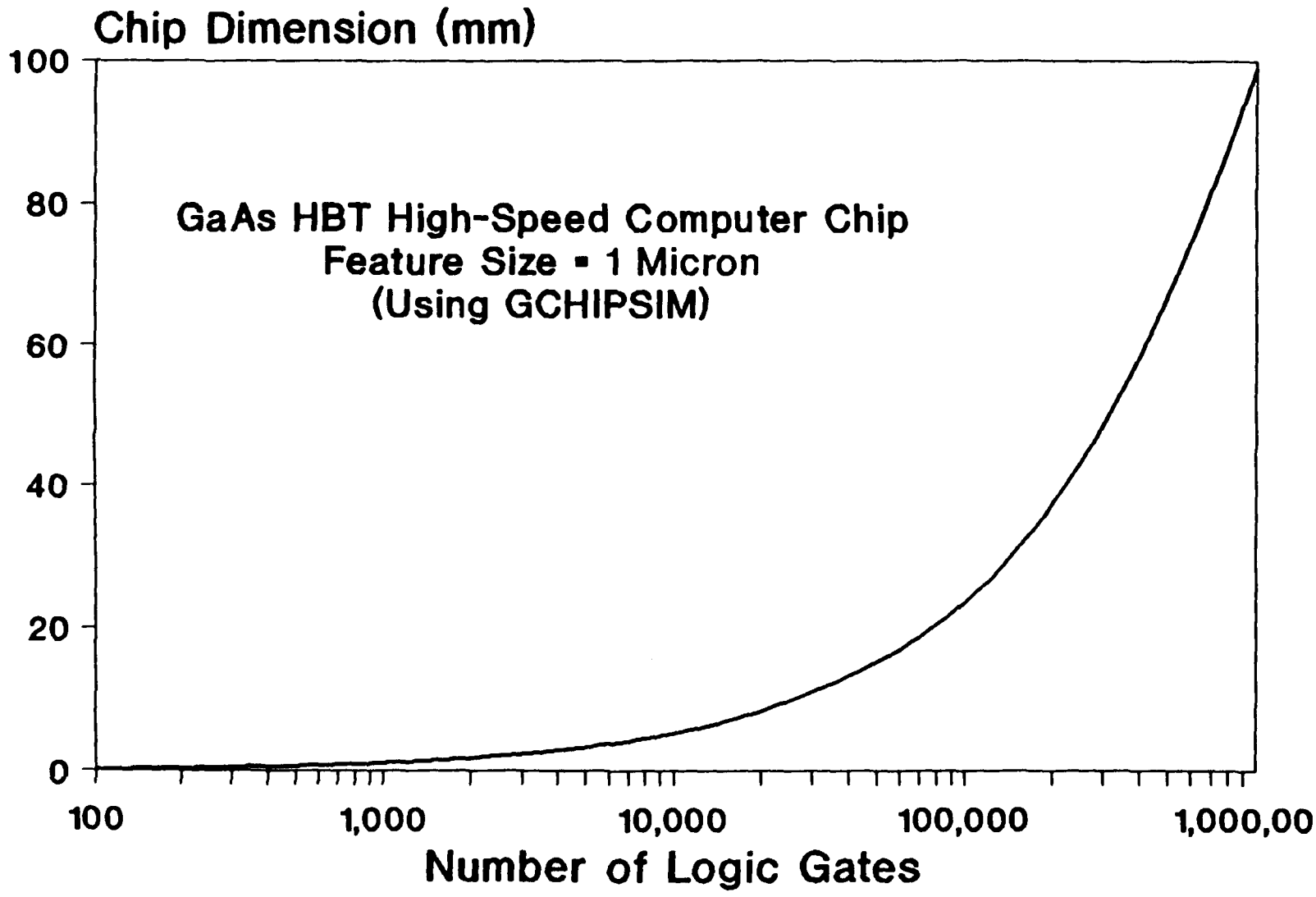

Chip Dimension (mm)
100
80
60
40
20
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 44

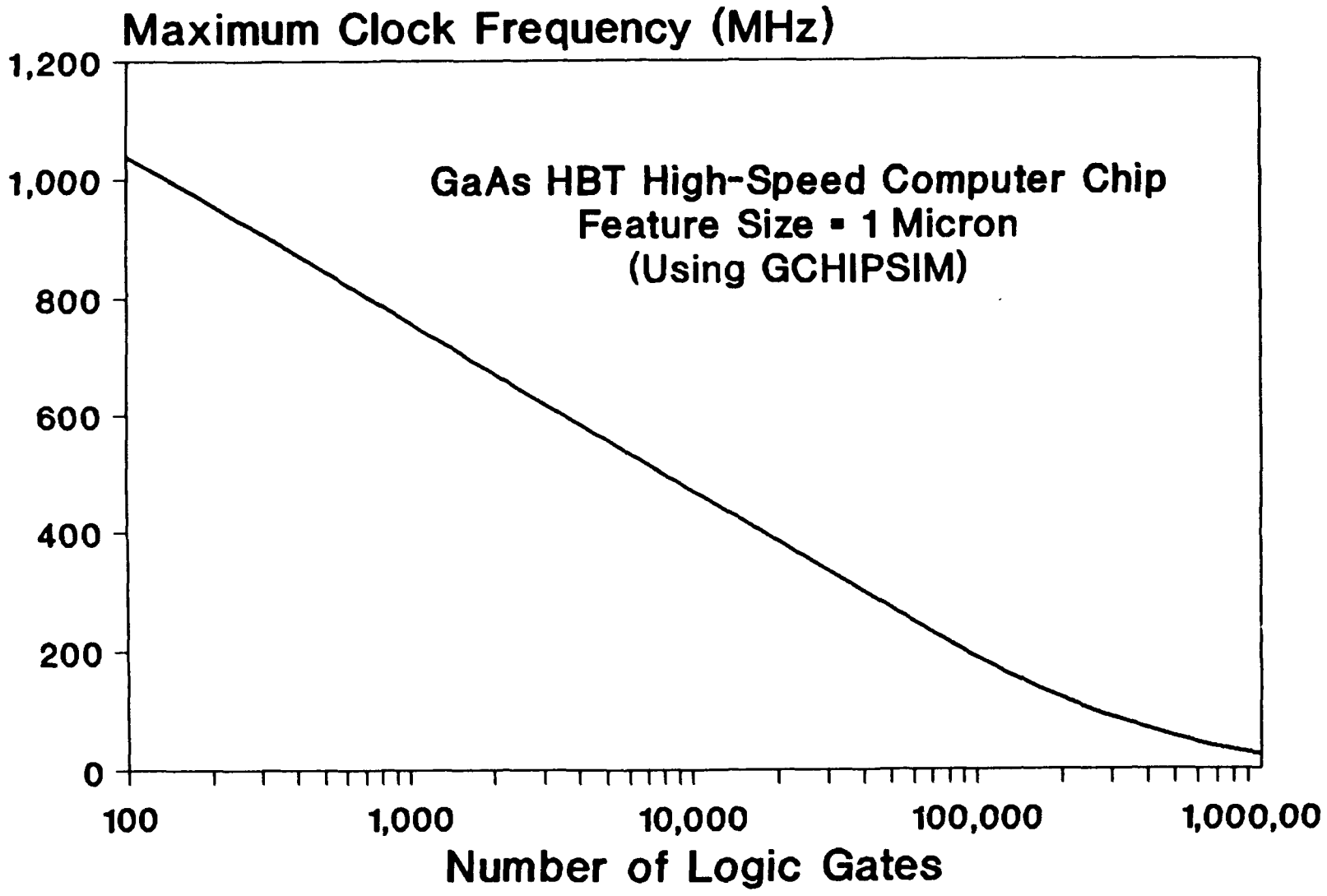

Maximum Clock Frequency (MHz)
1,200
1,000
800
600
400
200
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 45

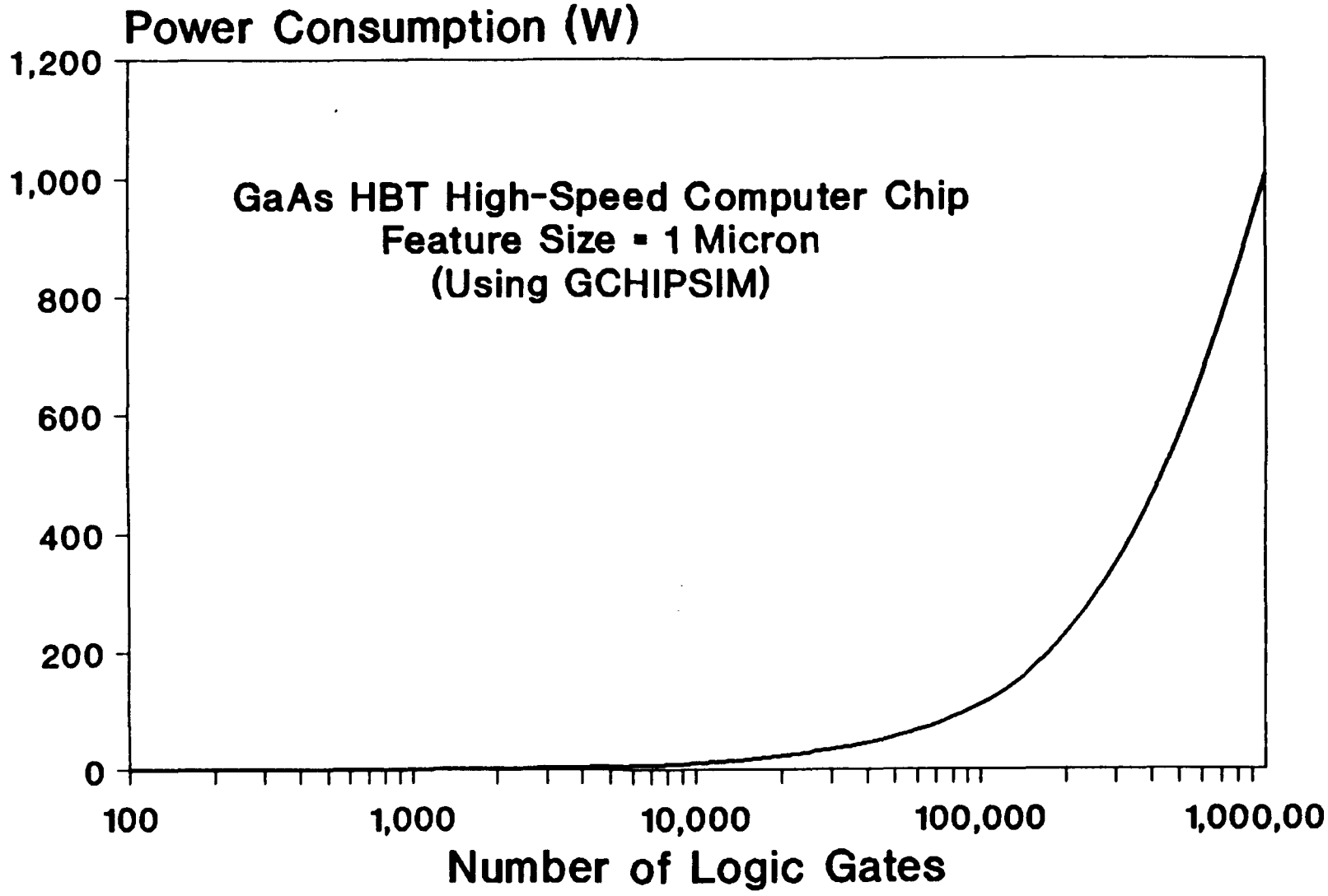

Power Consumption (W)
1,200
1,000
800
600
400
200
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 46

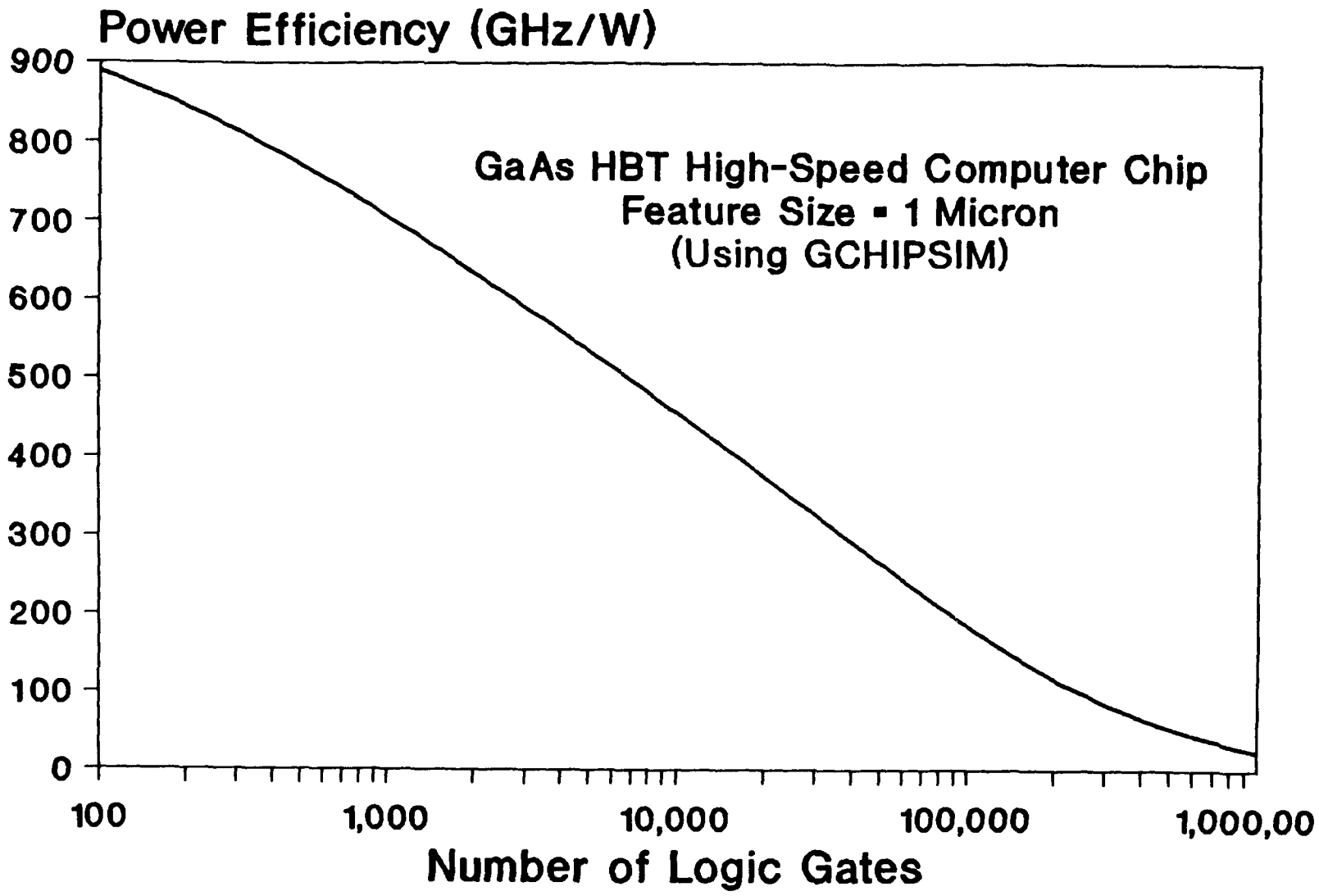

Power Efficiency (GHz/W)
900
800
700
600
500
400
300
200
100
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 47

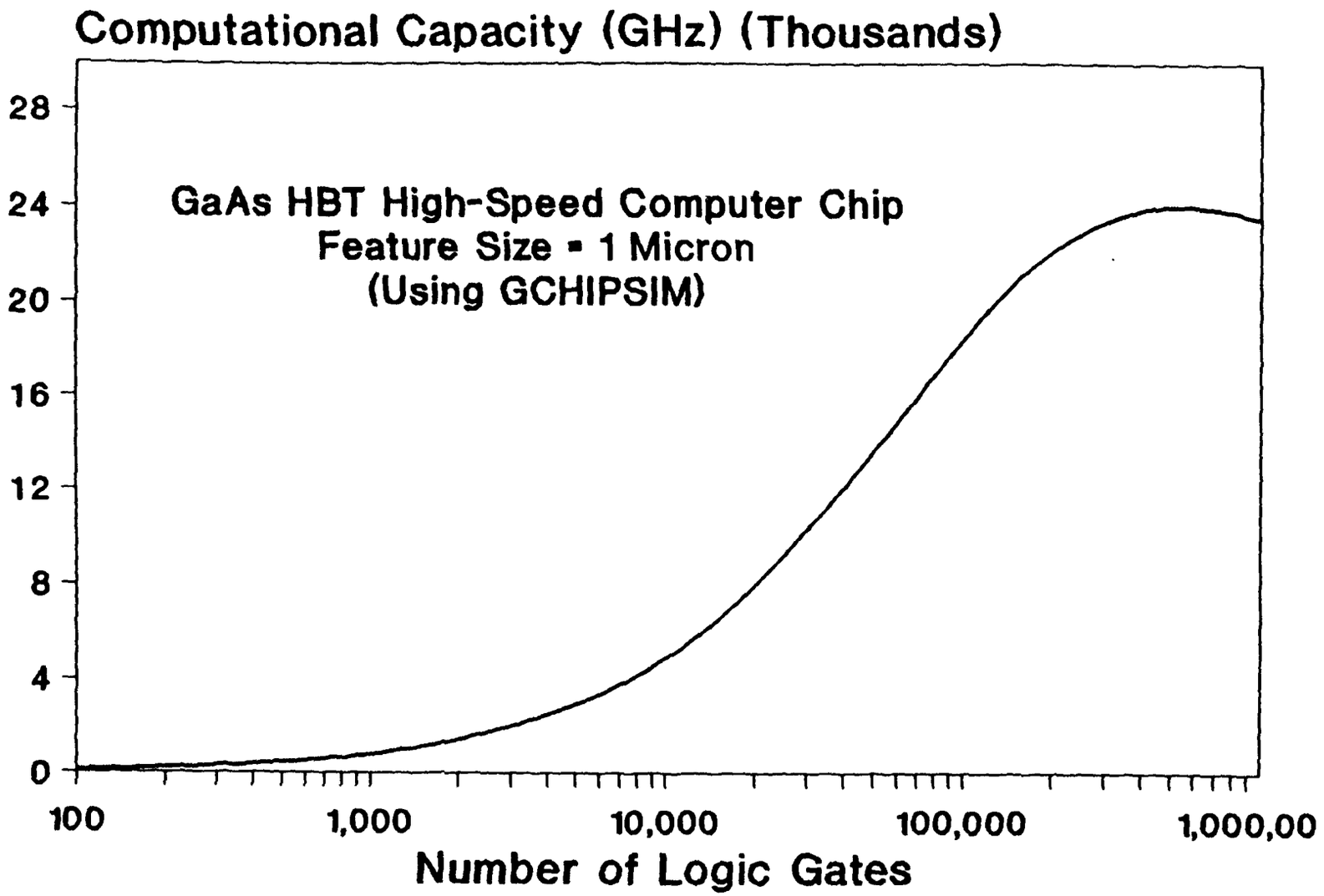

Computational Capacity (GHz) (Thousands)
28
24
20
16
12
8
4
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 48

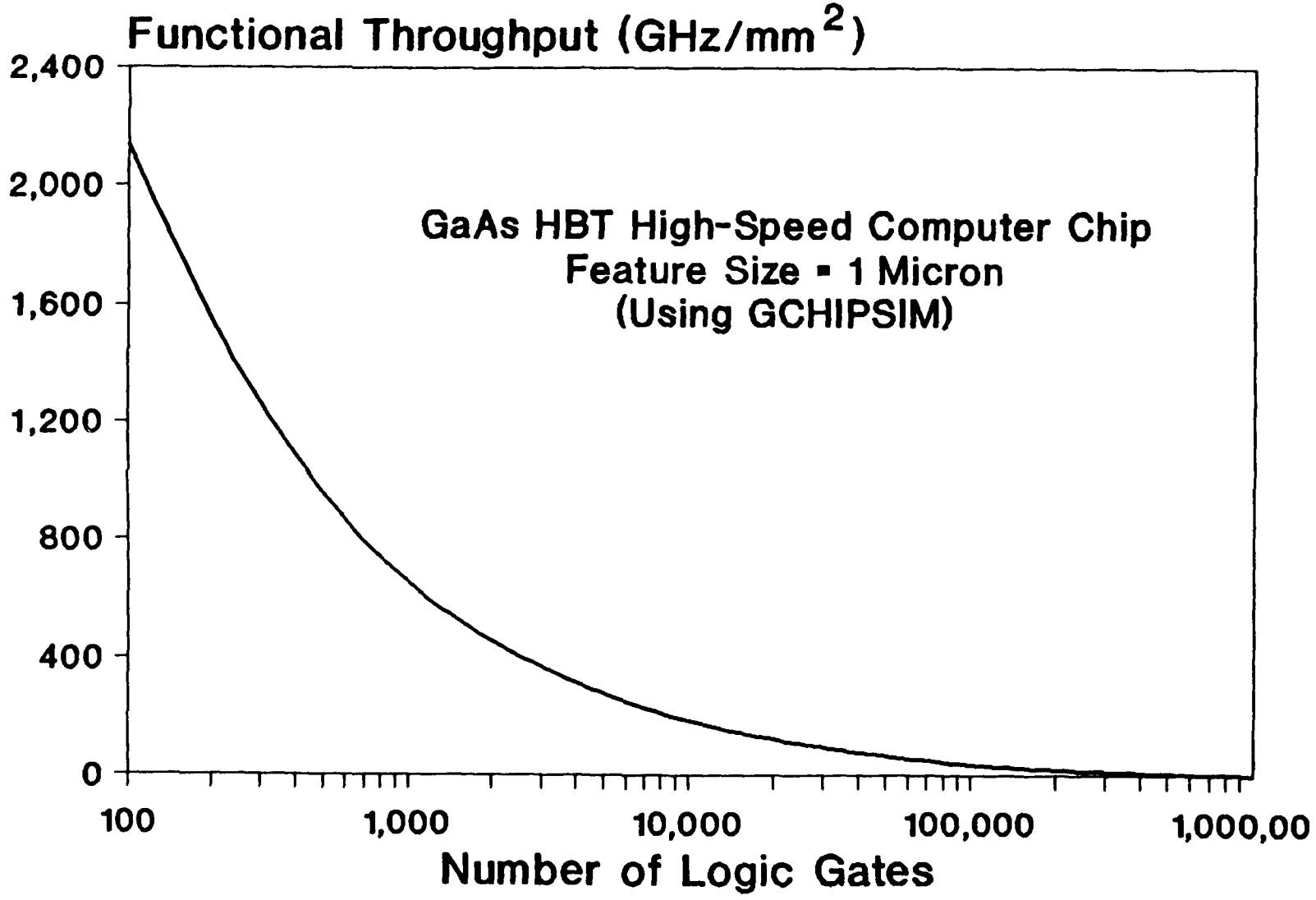

Functional Throughput (GHz/mm^2)
2,400
2,000
1,600
1,200
800
400
0
GaAs HBT High-Speed Computer Chip
Feature Size = 1 Micron
(Using GCHIPSIM)
100
1,000
10,000
100,000
1,000,00
Number of Logic Gates

FIGURE 49

7.2 PROGRAM LISTINGS

The listings of the programs NCHIPSIM, CCHIPSIM, BCHIPSIM and GCHIPSIM can be made available on request.

BIBLIOGRAPHY

1. Donath, W. E., "Placements and Average Interconnection Lengths of Computer Logic," *IEEE Trans. Circuits and Systems*, Vol. CAS-26, April 1979, pp. 272-277.

2. Landman, B. S. and Russo, R. L., "On a Pin versus Block Relationship for Partitions of Logic Graphs," *IEEE Trans. Computers*, Vol. C-20, Dec. 1971, pp. 1469-1479.

3. Price, J. E., "A New Look at Yield of Integrated Circuits," *Proceedings of IEEE*, Vol. 58, Aug. 1970, pp. 1290-1291.

4. Toong, H. D. and Gupta, A., "An Architectural Comparison of Contemporary 16-Bit Microprocessors," *IEEE Micro*, February 1981, pp. 26-38.

5. Toong, H. D. and Gupta, A., "An Architectural Comparison of Contemporary 32-Bit Microprocessors," *IEEE Micro*, February 1983, pp. 9-22.

6. "Cray Spins Cray-3 Out to Start-Up," *Electronic Engineering Times*, Issue 539, May 22, 1989, p. 1.

7. "Intel i486 Introduced; Integrates MMU, FPU," *Electronics News*, Vol. 35, No. 1754, April 17, 1989, p. 1.

8. Goel, A. K. and Huang, Y. R., "Parasitic Capacitances and Inductances for Multilevel Interconnections on the GaAs-based VLSICs," *Journal of Electromagnetic Waves and Applications*,

9. Mactaggard, R., "Typical Parameters for GaAs Chips," (Private Communication).

FINAL REPORT

STUDIES ON THE IMPROVEMENT OF THE ACCURACY OF ACOUSTIC EMISSION SOURCE LOCATION FOR SMART STRUCTURES APPLICATIONS

- Submitted to -

Universal Energy Systems, Inc.
4401 Dayton-Xenia Rd.
Dayton, OH 45432
P.O. No. S-210-11MG-077
Contract No. F49620-88-C-0053/SB 5881-0378
1990-91 USAF -UES FACULTY INITIATION RESEARCH PROGRAM

Sponsored by the
Air Force Office of Scientific Research
Conducted by the
Universal Energy Systems, Inc.

- Submitted by -

M. A. HAMSTAD (303) 871-3191
University of Denver
Department of Engineering
Denver, CO 80208

December 30, 1991

STUDIES ON THE IMPROVEMENT OF THE ACCURACY OF ACOUSTIC EMISSION SOURCE LOCATION FOR SMART STRUCTURES APPLICATIONS

- by -

Marvin A. Hamstad

ABSTRACT

An acoustic emission (AE) test platform was designed and tested after consideration of several alternatives. This test platform is intended for use in developing more accurate AE source location approaches useful for smart structures applications. The major design objectives of the test platform included: a) flat with frequency sensors in the far-field of crack related AE sources; b) a uniform thickness plate section containing a crack with sufficient lateral dimensions to remove early edge reflections of the acoustic waves; and c) ability to test the specimen in an existing acoustically quiet screw-driven test machine with hydraulic grips to minimize extraneous AE sources. In addition, a literature search identified two potential techniques to improve AE source location accuracy. Preliminary data for real AE was obtained with both techniques for the first time. Typical waveforms are shown for a small array sensor and for an in-line sensor configuration for crack opening AE events in aluminum. Recommendations for extension of this limited study are presented.

ACKNOWLEDGEMENTS

The excellent efforts of Dr. George P. Sendeckyj (Fatigue, Fracture and Reliability Group, Wright Laboratory) in making the arrangements for testing, certain specimens, and instrumentation at Wright Patterson AFB is gratefully acknowledged. Also the fine efforts of other members of this group are acknowledged as well. Also the work of Vikas Malik (graduate student, Department of Engineering, University of Denver) is acknowledged.

Finally, the support for this research by the Air Force Office of Scientific Research is acknowledged as well as the administrative support of Universal Energy Systems.

I. INTRODUCTION

The Fatigue, Fracture and Reliability Group of the Air Force Wright Laboratory is interested in the concept of smart structures. This terminology refers to the ability to nearly continuously sense conditions and states such as loads and structural integrity in aircraft structures [1]. The inputs of the sensors into properly programmed computers lead directly to corrective actions thereby bypassing scheduled nondestructive evaluation. Acoustic emission {AE} technology provides one approach to the network of sensors and computers necessary for smart structures.

AE technology is concerned with the monitoring of stress waves hich are generated by rapid local energy releases in solid materials. The damage generated (or related) AE signals are created by a variety of sources such as bond fracture, impacts, phase changes, crack growth, friction, inclusion-particle fracture, and other microdamage sources.

AE technology has a number of useful and potentially unique features applicable to smart structure applications. First, AE is a passive technique in that stress waves generated at damage sources throughout a structure propagate to the sensor locations. This feature limits the density of sensors required to monitor a structure. Second, AE is a whole volume damage sensing technique no less sensitive to interior damage compared to surface damage. Third, AE data from multiple sensors can be used to locate the spatial position of the damage sources within the structure. Fourth, the damage source mechanism may be potentially identified by use of advanced statistical concepts with waveform feature vectors. Fifth, AE is essentially a real time damage sensing technique. Sixth, AE is a microscopic characterization technique with high sensitivity. And seventh, AE is at its most fundamental level a measure of the damage response of a structure to stress.

Since an important goal of the smart structures concept is to bypass the usual nondestructive inspections, the application of AE would make its most valuable contribution if the extent of damage in a structure could be continuously defined by using the source location capabilities of AE technology. For example, with a metal structure in a fatigue environment, crack lengths could be measured by continuously following the position of the crack tips with AE source location data. Such an application of AE requires an extension of the technology as usually practiced in structures. Typical AE approaches are based on locating the vicinity of a crack or damage region rather than actual location of crack tips or the advancing perimeter of a damaged region. Thus, the smart structures application requires a significant increase in the accuracy of AE source location techniques in structures.

Examination of the accuracy of AE source location with conventional commercial AE instrumentation demonstrated that such an approach does not locate crack tips [2]. Further, it was shown in the same reference that an approach using waveform recorders with commercial resonant sensors results in only a small improvement in source location accuracy. Based upon the above results, it can be concluded that approaches which ignore source radiation field characteristics and specimen wave propagation features will not be capable of accurate source location.

On the other hand, accurate location has been demonstrated for waveform recorders by use of nearly flat with frequency sensors located very close to the AE sources. The keys to success with flat sensors are: 1) the velocity(s) of the specific wave(s) mode used to locate the source must be known; 2) the path from source to sensor of these wave modes must be known; and 3) the same phase point within the wave mode must be used. Consistent with such an approach, Scruby [3] and Johnson and Carlson [4] quote accuracy values on the order of ±500 μm and ± 100 μm

respectfully. These results were obtained using multiple sensors of the point contact type with waveform recorders. The best results, Johnson and Carlson [7], had a total of 8 sensors located on both surfaces of a plate nearly as close to the crack as was physically possible. The other quoted results were for a relatively thick compact fracture toughness specimen with sensors on the order of 3 cm from the crack. Thus, in both cases, with sufficient source amplitude, the first arrival of the direct path P wave (compressional wave) could be identified and used to obtain relative in phase arrival times.

Clearly, a smart structures AE application will generally have sensors at greater distances from the source to maintain a reasonable sensor density. In such a case the wave propagation aspects become considerably more complicated since the direct path compressional wave has insufficient amplitude. Assuming a range of plate thicknesses of 0.125 in. to 0.625 in. for sensor spacings of 8 to 20 in. or more in structural applications, sensor to source distances are some 13 to 160 times the plate thickness. Thus, the AE signals will be dominated by appropriate plate modes in the so-called far field. Since no literature results on location accuracy of real AE sources in plates with sensors in the far field could be found, a beginning was made to study this area. The results of a limited Research Initiation Program Mini-Grant study are given in this report.

II. OBJECTIVES OF THE RESEARCH EFFORT

This research had two objectives. The first objective was design, development, and qualification testing c an AE test platform which would be appropriate to study source location accuracy of real AE in a plate at source to sensor distances appropriate to smart structures AE applications. To meet this objective the transverse size of the

plate specimen was required to be such that source to sensor distances would put the sensors in the far-field of the source. Further, at the same time the plate size must place the sensors sufficiently removed from the plate edges or thickness change regions so that a sufficient portion of the AE signal could be observed without distortion due to reflections from these boundaries.

The second objective was to obtain initial multi-channel AE waveforms with relatively flat sensors applied to the test platform containing a sharpened crack under fatigue and overload testing. This objective involved exploring the theoretical and experimental wave propagation literature as related to AE signals in plates. The purpose of the literature search was to determine the best potential approaches for sensor types, positions, frequency ranges and analysis techniques to accurately determine the source location of real AE crack-based sources monitored in the far field. Then using the best approach(es) the waveforms obtained would be examined to determine the potential usefulness of the approach(es) for accurate source location.

III. AE TEST PLATFORM: DESIGN AND PRELIMINARY TESTING

Since one of the primary objectives of this research was to develop an AE test platform for use in a future extensive evaluation of various approaches to accurate AE source location using AE sensors in the far field, a paper design study was first made to identify the best design that would meet AE test requirements and minimize the cost of the specimens. The key constraints in this design study were: 1) a center-cracked plate of sufficient lateral dimensions to eliminate early acoustic wave reflections at sensors in the far field; 2) a specimen and gripping system which would generate little or no extraneous AE; 3) total specimen size such that it could be tested with existing hydraulic grips in the acoustically quiet screw-type Instron machine

available at Wright Laboratory; and 4) minimum cost to manufacture additional future samples.

Designs considered in the study included: 1) specimens machined out of solid plate stock; 2) specimens with adhesively bonded transition tabs; 3) specimens with bolted transition tabs; and 4) specimens with welded transition tabs. The bolted-type specimens which have been successfully used for non-AE type testing were eliminated due to the presence of extraneous noise sources at the bolt to hole interfaces as well as at the tab to plate interfaces. Welded specimens were eliminated due to the expected difficulties with warping and the expected brittle weld material which could lead to extraneous AE as well as tab failures during fatigue cycling for crack sharpening. The specimens chosen for testing were the specimen machined from solid stock and the specimen with adhesively bonded tab transition regions.

Since frictional and interface extraneous noise sources are not present in a solid sample, full-size samples with two different central plate section thicknesses were designed and fabricated from 2024-T351 aluminum. Figure 1 shows the dimensions and other pertinent information on this specimen with unusual geometry. Because the overall axial length of the specimen was limited by the test machine and the essential design constraint was to eliminate early acoustic reflections, a specially designed transition region between the 3.75 in. grip width and the 34 in. wide plate region was not used. Such a transition region would result in more uniform stresses in the plate section of the specimen, but at a penalty of a significant reduction in the length of the plate section in the axial direction. This condition would result in early acoustic reflections. Instead the design used the minimum length transition region sufficient to control stress concentrations as determined from reference 5. The resultant design with the selected location for the small array transducer (to be

discussed later) at a nominal 6 in. distance from one crack tip and a 30° angle from the crack plane gives an acoustic wave time window of about 50 μs from the arrival of the extensional mode wave at the plate wave extensional velocity [approximately 0.214 in/μs] until the arrival of the first reflection of this wave. The calculated window until the arrival of the first reflection of the dominant flexural mode was about 120 μs [based on a velocity of approximately 0.114 in/μs].

Since the adhesive tab design was unproven, both with respect to strength, fatigue life, and extraneous AE noise generation, a scaled-down sample was designed and tested to develop data with respect to these unknown aspects. Figure 2 shows the relevant dimensions of this 0.125 in. thick sample. This subscale design maintained the stress concentrations as well as the shear stresses in the adhesive bonds at the same levels present in a full-size adhesively-bonded sample. A special room temperature curable epoxy adhesive which does not contain a filler that would generate extraneous AE was obtained from Dexter/Hysol (EA 9330.4). This adhesive has excellent strength (4000 psi tensile lap shear) and was expected to have good fatigue properties although no direct fatigue data were available.

An initial evaluation of the extraneous noise characteristics of this sub-scale sample was carried out by proof testing the sample to 25,000 lbs. with AE monitoring. This proof test was carried out on an MTS machine with hydraulic grips at the University of Denver. This test machine is modified in that the servo valve is off-mounted from the actuator to control machine noise. An additional modification was necessary to control noise from the cycling of the pressure pump that raises the gripping pressure (above the 3,000 psi supply pressure) of the hydraulic grips with smooth inserts. Installation of needle valves in the lines to the grips allowed the grips to be pressurized and then valved off during the tests. This approach permitted

shutting off the grip pressure pump for the duration of the test.

The AE system used in this test (Physical Acoustics 3000/3104, μ-30 sensors bonded to the sample, 100-300 kHz, 54 dB threshold with 60 dB preamplifier gain) indicated hundreds of AE events starting at loads of only a few hundred pounds. Since these loads were well below those required, it was decided not to make a full size adhesive tab type specimen. This specimen was later notched (with a water jet) and tested in fatigue at Wright Laboratory. The adhesive bond failed in the tab region prior to the central starter notch growing to the intended size of 1.00 in. This failure occurred after about 9,000 cycles under a nominal ΔK of 20 Ksi$\sqrt{in}$ after reaching a crack length of 0.720 in. (initial notch size was 0.500 in.). It was thus concluded that the currently designed adhesive sample also did not have sufficient fatigue life for the necessary pre-cracking procedure. A different design which results in a more uniform distribution of shear stresses may possess sufficient fatigue life, but this design would not likely pass the extraneous noise requirement.

Since cost was also a consideration, it is of interest to note that the solid full-sized samples cost about $1200 each for materials and machining, while the adhesively-bonded subscale specimen cost about $500. With the additional cost expected for a full-size adhesive specimen it does not seem worth the risk of extra extraneous noise sources.

After machining the solid samples, each was strain gaged along the line shown in figure 1 and then loaded in tension. The purpose of this preliminary test was to allow an estimate of the stress intensity factor at the notch/crack tips during the fatigue pre-cracking process. By comparing the measured strain levels with those for a plate with uniform strain, an effective width of the samples was obtained. This effective width was used to calculate the approximate stress intensity factor. Table 1

shows the pre-cracking history (for both samples) used to extend the initial notch to the 2.00 in. crack size with an approximate maximum stress intensity factor of 10 Ksi√in, which is well below the K_{IC} value for this material of about 35 Ksi√in.

Prior to the AE tests of the load platform an additional test was necessary to check for extraneous grip noise. The previous work had shown with all the precautions taken that extraneous noise was not a difficulty with smooth hydraulic grip inserts [2]. Since the two test platforms had different thicknesses in the grip region compared to the previous test specimens, it was necessary to use different grip inserts. These different inserts were serrated rather than smooth, and smooth inserts were not available in a reasonable time or at reasonable prices since they would have to be made as a special order in England where the hydraulic grips are manufactured. Hence, using the AE system and sensors previously described in reference 2, a dummy aluminum sample was tested for extraneous noise in load cycling (0 to 22,000 lbs) above the loads planned for the solid test platform. These tests which were carried out at the same maximum sensitivity (limited by electronic noise) of 42dB threshold out of a 60dB gain preamplifier showed only a few events, as documented in table 2. It is speculated that the serrated grip inserts do not contribute additional noise as a result of the procedure of applying 3,000 psi hydraulic pressure to the grips and then reducing the pressure to 2,700 psi prior to applying axial load to the specimen.

IV. SELECTION OF AE APPROACH AND PRELIMINARY TESTING

Since the limited nature of this research project dictated that totally new approaches could not be developed for accurate AE source location, the best that could be done was to extract from the published literature previously developed approaches which incorporated results from wave propagation theory for accurate location using

AE sensors in the far-field of the source. An extensive search of the AE literature as well as discussions with others in the AE research community during the time frame of January to March of 1991 identified two potential approaches.

Sachse and Sancar [6] were issued a patent in 1986 based on an approach that used a closely spaced array of four small diameter approximately flat with frequency (velocity sensitive) sensors. This approach is only documented in the open literature by the patent which lacks the type of discussion usually found in a typical published paper or report. Only after extensive efforts was a copy of a hand-written report, upon which the patent was based, obtained in late October 1991 [7]. The patent describes an accurate AE location approach based upon visual selection of the same waveform features in the waveform from each of the four sensors in the array. For each of two separate features, the group velocity is calculated, and then using these two group velocities and the difference in arrival times of the selected features, the range and direction to the source are calculated. It is important to note that the experimental work upon which the patent was based included only simulated AE sources (lead breaks) applied perpendicular to the plane of a plate. No testing had been done with real AE where typically inplane AE sources are present. In fact the theoretical far-field solution for an inplane AE source is not in the published literature to correspond to the solution available for the out-of-plane case [8].

Ziola and Gorman [9] have recently been developing a different approach to accurate AE far-field location in thin plates. Their approach is based on use of a cross-correlation technique to determine the proper phase points within AE waveforms from conventionally-spaced approximately flat with frequency sensors. The approach and analysis is based upon applying classical plate theory (thin plates) which includes only the extensional mode and the flexural mode of the more complex Lamb theory. As

with the above small array technique this cross-correlation approach has to date only been demonstrated for out-of-plane simulated AE (pencil lead breaks) rather than for real inplane AE sources.

It is important to note that both of the above approaches allow for dispersion of the acoustic waves and thus they overcome a key reason for the inaccuracy of conventional AE source location approaches. At present both approaches also do not account for edge reflections. Thus they are consistent with the current design objectives for the AE test platform. The fact that neither approach has been tested with real AE made both approaches good candidates for the testing aspect of this research.

Since only one approach fit the budget constraints of the project, the small array transducer (SAT) technique was chosen since this approach may be especially suitable for smart structures applications. The reason that this technique is particularly suitable is that, since all four sensors are in close proximity to each other, they lend themselves to local processing of data. Thus after local processing only the results need to be sent on to more central computing stations. With the alternate approach the raw data from each sensor must be sent on, thereby greatly increasing the amount of data requiring transport over significant distances. Further, the SAT technique is expected to be less sensitive to the effects of the radiation pattern differences from one source to the next since all four sensors are at about the same angle to a given source.

Later in the project it was decided to invite Dr. Michael Gorman to participate in the testing at Wright Laboratories, since his participation would not add to the cost because he had available support from the Astronautics Laboratory at Edwards Air Force Base, CA. Since the AE test platform is large and has two sides, it can support

up to four simultaneous AE approaches without "congestion" of sensors. Preliminary results of Dr. Gorman's participation are included in Appendix I to this report.

Since the SAT technique depends on a particular sensor called a pinducer (model VP-1093 available from Valpey-Fisher), it was necessary to determine the relative output of this sensor compared to those used previously in order to make sure that the expected loss of sensitivity with a flat sensor would not result in a total loss of capability to detect the AE from the 2024-T351 aluminum. Table 3 lists relative decibel outputs of the pinducer (velocity sensitive) as well as two other relatively flat displacement sensors compared to the output of the resonant sensor for which the typical amplitudes for AE crack sources in 2024-T351 are known from previous work [2]. These output levels were obtained in the far-field using top and midplane edge lead breaks as the source. The relative outputs were determined for the dominant wave groups, namely the first extensional mode arrival and the first flexural mode peak. Since the previous work indicated a range of amplitudes from 42 - 80 dB with the μ-30 sensor, the pinducer being about 25 dB less "sensitive" was still expected to detect AE signals. But, the number detected would be considerably less and hence effort is needed to develop higher output flat sensors and/or less noisy preamplifiers.

Since no published paper (other than the patent) was available on the SAT approach, the method described in the patent was first tried using data constructed by assuming an SAT at a nominal distance of 6.00 in. from an assumed source position. Using the velocities of two wave groups expected to be dominant for a real source with sensors in the far-field, the arrival times were calculated for a source on the same side of a plate as the SAT position. Figure 3 shows the geometrical experimental configuration. The propagation distances were calculated to eight decimals based on the location of the center of each pinducer. Then using 0.214 in/μs and 0.114 in/μs

as the velocities of the dominant groups in the lowest symmetric and antisymmetric branches the propagation times were calculated to eight decimals. From these propagation times the appropriate differences in arrival times were calculated as well. To simulate in part the effects of different finite digitization rates the arrival time differences were rounded to the appropriate number of digits. Appendix II gives the equations used from the patent to calculate the group velocities, range, and direction to the source. Table 4 shows the values and the percentage errors based on the known values from which the propagation values were calculated. Since the ratio of the range to the source relative to the maximum array dimension was on the order of 6, it was necessary to apply a correction term (see Appendix II) to the approach indicated in the patent when the range is much greater than the maximum array dimension. Clearly, from table 4 a digitization rate of at least 0.1 μs is required for accurate results in this ideal numerical case.

V. AE TEST APPROACH LARGE PLATES

After the center cracks were sharpened out to a 2.00 in. length by cyclic fatigue at about 0.5 Hz, each specimen was mounted in the 100 kip Instron for testing with AE. Two sensor arrays were used for the four pinducers. First, tests were made with the SAT configuration (at 30° to the crack plane, see figure 3) suitable for source location following the patent. Second, additional tests were made with an in-line configuration with all four pinducers spaced 2.00 in. apart along an imaginary line emanating from one crack tip at 30° to the crack plane (see figure 4). The 30° position was chosen to provide for observation of significant amplitudes of bulk longitudinal and shear radiation for a crack plane AE source. The sensors were held in position with an aluminum cantilever bar that was clamped to the thick shoulder

of the specimen. The pinducers were bonded with adhesive (Super Glue) to the inside diameter of a hollow rod of micarta as shown schematically in figure 5. Then the micarta rod was inserted into close tolerance slip fit holes in the cantilevered bar until the sensor came into contact with the plate. A tapered wedge was then inserted into the slots at the top of the micarta to expand it to hold it firmly in place.

The four sensors used were pinducers with coaxial pinducer cables (Valpey-Fisher models VP-1093 and VPC-4, respectfully). Modified preamplifiers (Physical Acoustics model 1220A) at 60 dB gain with a nominal bandpass of 0.035-1.0 MHz (3 dB down points) were used powered by the nominal 28 volts from a 24 channel Spartan (Physical Acoustics Corp.) expansion box. The expansion box was used only to supply the power to the preamplifiers and to properly separate the AE from the dc power. For most of the tests the internal power supply of the Spartan was not used. Instead several 6-volt "lantern" batteries were connected in series and then to the distribution board of the Spartan. This approach was used to try to decrease background electronic noise. For data recording a four channel waveform recorder (Nicolet model 440, 12 bits, 0.1 μs digitization interval) was connected by electrical tees at the Spartan distribution board. All four channels were triggered simultaneously to save typically 4000 data points (including pre-trigger information). These data were stored onto 3-1/4 inch standard disks.

Some electrical noise originating from an unknown unshielded component in the Instron was discovered at an approximate burst frequency of 10 kHz. A braided cable was connected between the plate specimen and the Spartan box to eliminate the noise. In order for this approach to be successful the original couplant selected for the pinducers had to be changed. The pinducers do not have an insulating wear plate where they come into contact with the test specimen (as do all commercial-type AE

sensors). The original couplant Apiezon M did not maintain sufficient electrical insulation even though it is sufficiently viscous that it does not flow perceptibly down a vertical surface under the influence of gravity. Substitution of a very stiff couplant wax called Tacki-wax (amorphous wax with a high melting point, Cenco) provided sufficient insulation combined with the ground cable to eliminate the problem. Testing with 0.3 mm pencil lead indicated no loss of sensitivity with the new couplant.

The original plan was to trigger the transient recorder with one pinducer channel. This approach did not work well because the level of the AE waveform was not significantly above the preamplifier electronic noise peak levels. Hence to eliminate false triggers and at the same time to increase the number of AE events captured, a fifth AE channel was added with a significantly more sensitive resonant type sensor (Physical Acoustics model R-15). The signal from this sensor located on the opposite half of the plate (relative to the crack plane) compared to the SAT was used as an input to trigger all four pinducer channels simultaneously. This trigger sensor was coupled with the same couplant and used a standard preamplifier (Physical Acoustics 1220A, 200-400KHz, 60 dB).

To gather the sample waveforms from real AE sources with approximately flat AE sensors located in the far-field of a uniform thickness plate of sufficient size to eliminate early edge reflections the specially designed AE platform was loaded in displacement control between zero and approximately the maximum load seen at the end of the precracking. Table 5 gives the loading conditions for both specimen thicknesses. Prior to and after each load cycle lead breaks [Pentel, 0.3 mm, 2H] were made at one or both crack tips for several purposes: 1) to verify proper acoustic coupling of each sensor; 2) to check for proper operation of the electronics; and 3) to provide data for a source at the crack tips with a high signal to noise ratio. In some

cases the lead was broken at the crack tips on the side opposite to the location of the pinducers and trigger sensor. It should be noted that this out-of-plane source would not be expected to partition energy in the plate modes in the same fashion as an in-plane source, but we were unable to devise an appropriate artificial in-plane source at the crack tips. During this cyclic testing the expected AE source was that due to crack opening causing fracture of the cold welds as observed previously in reference 2. To increase the number of events, significant rest periods at zero load were incorporated in the test schedule, since the previous results had shown the usefulness of such an approach [2]. Since it was desirable to limit sets of recorded waveforms to those with reasonable signal to noise ratios, interrupted cycling was used. Hence, when appropriate waveforms appeared on the Nicolet display the crosshead was stopped. The pause gave time to record the waveforms to disk and to note the load at which the AE source event occurred. The test was then resumed until the next waveforms appropriate for storage appeared. After a number of such fatigue cycles, the specimen was given some overload cycles to attempt to obtain AE waveforms from inclusion particle fracture in the crack tip plastic zone. Table 5 also gives the maximum overload levels for each specimen. The overload cycles were applied with manual control of the Instron in 20 to 50 lb load increments. This procedure was used since the previous work [2] had indicated relatively high rates of AE during the overload. Hence, if this overloading was done at a fixed crosshead rate at most one set of waveforms would be recorded. Since no overload events were recorded, a final set of overload cyclic tests were carried out on the 1/8 in. thick sample with no interruptions or manual control. These cycles were made at an increased crosshead rate as well so that the cyclic rate was on the order of 0.08 Hz. Again, no overload events were recorded.

Finally after the cyclic loading and overload tests were completed, a series of pencil lead breaks (Pentel, 0.3 mm, 2H) were carried out on one edge of each sample. The lead was broken near the top edge (test platforms were horizontally supported for these tests) of the plates at a point in the crack plane. Two pinducers were coupled to the top surface with Apiezon M. One pinducer was placed at a fixed position near the plate edge to act as a fixed trigger. The second pinducer was moved after each lead break in increments of 1/8 or 1/4 inch through a range of distances from about 3 inches to 9 inches from the source position. These tests were done to characterize the features of the typical wave transient propagation in the test samples. The edge break was used to simulate the in-plane nature of the real AE sources.

VI. RESULTS AND DISCUSSION

The success of the AE test platform in providing a test specimen where advanced techniques for more accurate source location of real AE sources can be studied is demonstrated in figure 6. This figure shows large amplitude high fidelity waveforms from the crack opening source for the SAT configuration on both plate thicknesses. These waveforms allow identification of both the extensional and flexural modes (see labels in figure 6) which are expected from either the simple plate wave theory or the more complex Lamb wave analysis. Further the required time interval of observation is completely free of the complications of either free edge reflections or reflections from the change in plate thickness at the shoulders of the specimen. Hence, available theoretical results can be applied to analyze these results. And at some time in the future when the solution for normal surface displacements in the far-field due to an inplane AE type source is presented, even more detailed analysis can be done. As previously noted, to date only the out-of-plane AE type source far-field

solution has been published [8].

It should be pointed out that this successful test fits the other constraints for useful AE testing such as no significant extraneous AE noise sources. The key factors which lead to lack of noise are use of a solid specimen, center crack with hydraulic grips which can be loaded prior to axial load application, and a relatively quiet screw-driven test machine. It was also important to demonstrate with the test platform that real AE sources are detectable with the less sensitive flat with frequency sensors. Data from this type of sensor are essential for use with theoretical results. Certainly as can be concluded from figure 6, more work needs to be done to improve the signal to electronic noise ratio. Such improvements will not only allow more fundamental analysis techniques to be applied but will also increase the number of AE signals detected thereby increasing the detectibility of smaller and smaller AE sources. To increase signal to noise ratio, both higher output flat sensors and/or less noisy preamplifiers need to be developed.

The test platform also meets cost constraints in the sense that a single specimen can be used for different source location approaches since the crack opening source is repeatable. Further, since the crack tips can be resharpened by additional fatigue cycling, even crack-tip overloads can be used repeatedly to generate AE when such a source is operative.

A further advantage of the test platform is that more than one experimenter can use it at the same time since both plate surfaces are accessible. And when the SAT approach is used, four separate quadrants could be used at one time on one side of the sample. Also, since both sides of the sample are accessible, studies on the differences in the partitioning of energy in the various wave modes can be made on two separate sides for the same source. Further, essentially 360° around the crack tip

can also be used so that radiation pattern effects in the far-field can be studied potentially for two separate source mechanisms (asperity cold weld fracture and inclusion particle fracture).

A potential weakness of the sample as currently designed is that the stress field in the uniform thickness plate portion of the sample is not uniform across the width of the sample. Figure 7 shows the axial strain versus position for one quadrant of the sample based on a finite element analysis for loading without a center crack. Clearly, the strains are higher in the center of the specimen in line with the loading tabs of the sample. For this initial design the potential for this effect was ignored as has already been discussed. As will be discussed later in this report, there are some reasons why a subsequent design of the test platform might include some transition zone provisions to meet other experimental goals.

As a first step in a more detailed examination of some of the waveforms obtained in this experiment, we consider figure 8. This figure for a crack opening event shows with an in-line configuration clear evidence of two AE events that occurred in close proximity in time. At about 10.8 μs after the arrival of the first event, the characteristic waveform of a second event appears. Clearly without the present experimental test platform such an observation would not be possible since in a more normal size test specimen edge reflections would compromise the ability to distinguish the second event in close proximity in time. And further, with a traditional resonant-type sensor, the waveform would be distorted. The distortion would make recognition of the second event difficult. It is of interest to consider the potential source location of the first and second events. Since from an AE wave propagation viewpoint the two crack tips are approximately in the near field, it is of interest to consider the times for bulk waves to travel from one tip to the other. This

time will give an indication as to whether the stress redistribution from the first event contributed to the generation of the second event. At a bulk p-wave velocity in aluminum at about 0.250 in/μs the 2.0 inch distance represents a time of about 8.0 μs, while the bulk s-wave velocity of about 0.120 in/μs represents a time of about 16.6 μs. Hence, it is possible that the local stress redistribution from one AE source event can trigger another event. Certainly this idea is open for future investigation, particularly if a sensor/preamplifier combination with increased sensitivity can be obtained.

Next we turn to an examination of the AE waveforms from the point of view of use of such waveforms for the purposes of accurate source location by the SAT configuration approach. The first aspect to consider is shown in figure 9 which is more representative of the signal-to-noise ratio of most of the crack opening AE waveforms at the SAT. Clearly preamplifier noise has a significant effect on the signal and will make more difficult the identification of the same waveform feature in all four sensors. Further, it is clear in this figure and also in figure 6 that it is almost impossible to clearly distinguish any features in the waveform prior to the arrival of the flexural mode. This observation is made more forcefully if we compare figures 6 and 9 with figure 10 which represents a lead break at the crack tip nearest to the SAT position and the in-line configuration. In this latter figure, both the extensional and flexural modes are easy to identify. There seems to be two potential reasons for the almost total lack of extensional signal in the real AE waveforms. First, the low signal to noise ratio makes it difficult to see the extensional mode. Second, it seems that the real AE signals from the present crack opening source deposit a dominant amount of energy in the flexural mode. For example, some of the largest amplitude waveforms where some of the extensional mode can be identified in the 1/4 inch plate have a typical ratio of flexural peak amplitude to extensional peak amplitude on the order of 15-20:1.

During the lead break experiments on the edges of the 1/4 inch plate, it was determined that the typical ratio was about 10-12:1 for top or bottom edge breaks and about 4:1 for the center of edge breaks, (both at about the same source to sensor distance as for the real AE). Since as Gorman and Prosser [10] have indicated, a center break deposits a greater portion of the energy in the extensional mode, the above results indicate that the crack opening source observed in these experiments was dominated by a source mechanism that deposited more energy into the flexural mode (as observed with vertical sensitivity sensors). In order to explain the reason for the dominance of the flexural source, it is worthwhile to consider an observation which was made during the fatigue precracking of the specimen. While fatiguing the samples there was a noticeable flexing motion of the plate which is consistent with the fact that the central portion of the plate experienced higher axial strains than the outer portions of the plate. Thus it is speculated that this flexing motion led to the crack opening source being dominated by fracturing of cold welds near the outer edges of the plate. To resolve this question further would require more detailed studies (beyond the scope of this initial study) of the elastic deformation of the sample with a center crack.

The key significance of the loss of extensional signal is that the demonstrated studies in the handwritten report by Sancar [7] specifically used extensional waveform features to determine one group velocity. In particular the very first extensional peak in the waveform is very easy to identify and use for a group velocity determination, since this plate wave (nondispersive) represents both low frequencies (i.e. long wavelengths compared to the pinducer element diameter and the maximum diameter of the SAT) and the displacement is always in the same direction as the wave propagates. The other feature which Sancar [7] typically used was the initial part of

the flexural mode. For this part of the waveform the phase point (feature) is more difficult to identify, especially in the presence of noise since the frequencies are higher (i.e. short wavelengths compared to maximum diameter of the SAT) and the sign of the displacement changes from tension to compression and back again with increasing propagation distance. Further, the shape changes with distance since the flexural mode is in general dispersive. Thus, without the extensional mode and difficulty in determining constant phase points for the flexural mode it was not possible to try the SAT approach to calculate the source range and direction. With additional work beyond the scope of this limited study, it is expected that constant phase points could be extracted from the flexural mode.

To begin to develop a rational way to select common features in the flexural mode a more detailed examination of the determination of group velocities was made using the top edge leadbreak data. The leadbreak data were chosen for this examination since they are free of the electronic noise. Figure 11 shows relative arrival time of the plate wave as a function of distance for the 1/8 inch thick plate. The slope gives the group velocity which compares well with expected values. The fact that most data points are right on or close to the line indicates the ease with which this is done and the reason why it is desirable to have sufficient sensitivity to detect the extensional mode for real AE. Figure 12 shows similar results for the flexural mode using the zero crossing with the greatest slope for the feature in the peak region. As can clearly be seen the slope gives a reasonably accurate velocity for the dominant flexural group that travels at approximately the shear velocity, but the fit is poorer, even over short distances on the order of the SAT maximum dimension. This result indicates that significant errors would occur in determining the correct time differences for use with the SAT approach where the close spacing of the sensors

requires very high accuracy of the measured time differences for accurate calculations of source location.

Significant numbers of AE events were observed for the crack opening source. These events occurred over a limited range of loads consistent with a fully closed crack at low loads and a fully open crack at higher loads with the events occurring at loads in between. Not as many events were observed as with the previous study [2]. But with the increased specimen size and more importantly the less sensitive sensors, it was expected to observe fewer events.

Virtually no events were observed in the overload tests which was contrary to the previous observations [2]. To investigate why the inclusion fracture AE source was not observed with the large samples, dogbone tensile samples were fabricated from both the previous 2024-T351 material and the current material. These samples were tested in the usual way for a uniform tensile sample at a constant crosshead rate of 0.050 in/min using smooth hydraulic grips, a resonant μ-30 sensor (Physical Acoustics), a bandpass of 100-300 KHz, and root-mean-square (RMS) voltmeters [Hewlett-Packard 3400A]. Figure 13 shows that the results were essentially the same from both material samples. Hence, both samples do have the same inclusion fracture source typical of the 2024 material. Thus we must seek another explanation for the lack of overload emission.

Another potential explanation is the change in the stress field in the large sample. Since, as figure 7 shows, the axial stress is not uniform, significant transverse stresses will also be present. Hence, the stress field at the crack tip will not be the same as for the uniform axial stress case. And the plastic zone size and shape could be expected to change as well. Without detailed studies (beyond the scope of this work) it is not possible to clearly determine if this change in stress field is the reason.

VII. RECOMMENDATIONS FOR FUTURE WORK

Based on the results of this study, four primary recommendations for future work are presented. First, effort should be invested in the development of flat with frequency sensor/pre-amplifier designs which have up to an order of magnitude better sensitivity than the currently available pinducers. Second, an additional AE test platform should be designed which will provide for a more uniform axial stress distribution in the plate section. Then specimens would be available which emphasize both extensional and flexural modes. It is likely that both specimens would be needed since real AE applications can result in both of these cases. Third, for the SAT approach additional effort should be made to develop a rational and automatic approach to identify the same feature point for the flexural mode. This work could start with Sancar's report [7], since it was obtained too late to be of much use in this initial study. Fourth, a theoretical far-field solution needs to be completed for an in-plane transient AE source.

VII. CONCLUSIONS

1) An AE test platform which meets requirements for generation and observation of far-field real AE waveforms with high fidelity without edge reflections or extraneous noise in a plate was successfully designed and tested for use in improvement of source location accuracy for smart structures applications.

2) This test platform concept provides for potentially two different AE sources, is reusable so that the cost can be spread over several different

test approaches, and can support more than one experimental approach at the same time.

3) A full test of the SAT approach could not be carried out due to the current difficulties of lack of extensional mode, poor signal-to-noise ratio, and difficulty in selection of the same phase point in the flexural mode.

4) Evidence was demonstrated that real AE source events can occur in close time proximity to each other.

REFERENCES

1. Sendeckyj, G.P., and C.A. Paul, "Some Smart Structures Concepts," Fiber Optic Smart Structures and Skins II, SPIE Proceedings, Vol. 1170, pp 2-10, 1989.

2. Hamstad, M.A., "Location of Crack Tips by Acoustic Emission for Application to Smart Structures," Final Report 1990 USAF-UES Summer Faculty Research Program, Universal Energy Systems, Dayton, OH, Contract No. F49620-85-C-0053, 1990.

3. Scruby, C.B., "Quantitative Acoustic Emission Techniques", in Research Techniques in Nondestructive Testing, Vol. VIII, edited by R.S. Sharpe, Academic Press, London, pp 141-210, 1985.

4. Johnson, J.A., and H.M. Carlson, "Applications of an Acoustic-Emission Data-Acquisitive Workstation: II", Proceedings of Review of Progress in Qualitative NDE 9, to be published, Plenum Press, New York, 1990, QNDE Conference 1989.

5. Kumagai, Kazuo, and Heihachi Shimada, "The Stress Concentration Factor Produced by a Projection Under Tensile Load," Bulletin JSME, Vol. II, No. 47, pp 739-745, 1968.

6. Sachse, Wolfgang H. and Selcuk Sancar, "Acoustic Emission Source Location on Plate-Like Structures Using a Small Array of Transducers," United States Patent No. 4,592,034, May 27, 1986.

7. Sancar, Selcuk, "Acoustic Emission Source Location and Groupe Velocity Determination on Plate-Like Structures Using a Small Array of Transducers," Personal communication, Oct. 1991.

8. Weaver, R.L. and Yih-Hsing Pao, "Axisymmetric Elastic Waves Excited by a Point Source in a Plate," J. of Appl. Mech., Vol. 49, pp 821-836, 1982.

9. Ziola, S.M. and M.R. Gorman, "Source Location in Thin Plates Using Cross-Correlation," J. Acoust. Soc. Am., Vol. 29, (5), 1991.

10. Gorman, M.R. and W.H. Prosser, "AE Source Orientation by Plate Wave Analysis," submitted to J. of Acoustic Emission, 1991.

Table 1. Pre-cracking History for Both Specimens with Initial Crack Length 1.00 Inch.

a. 1/4 Inch Thickness

Maximum Load, lbs	Number of Cycles	Finishing Crack Length (2a), in.
34,000	14,000	1.013
31,000	500	1.038
25,000	1,000	1.060
23,000	2,000	1.104
22,000 +	15,000	1.203
21,500	15,000	1.324
20,700	16,000	1.418
19,900	15,000	1.515
19,200	15,000	1.612
18,600	15,000	1.719
18,000	15,000	1.829
17,500 *	12,000	1.910
17,000	13,000	2.001

\+ Minimum load 10% of maximum
* Minimum load 1% of maximum

b. 1/8 Inch Thickness

Maximum Load *, lbs	Number of Cycles	Finishing Crack Length (2a), in.
19,400	500	1.01
17,000	284	1.02
15,500	417	1.0#
14,000	440	1.04
13,000	500	1.05
12,470	2,500	1.10
12,050	5,000	1.20
11,500	5,000	1.30
11,100	5,000	1.40
10,700	5,000	1.50
10,400	5,000	1.60
10,100	5,000	1.70
9,800	5,000	1.80
9,140	5,000	2.00

Table 2. Extraneous Noise Events from use of Serrated Hydraulic Grip Inserts with 60 dB Preamplifier, 42 dB Threshold, 200-400 kHz, μ-30 Sensor, Loads 0 to 22,000 lbs.

Test	Total Number of Events
Gripped, then 1st Cycle	15
2nd Cycle	9
3rd Cycle	5
Regripped, then 1st Cycle	8
2nd Cycle	3

Table 3. Relative Outputs from Different AE Sensors 6.0 inch from Leadbreaks on Edge of 0.313 inch Aluminum Plate for 35 KHz - 1.0 MHz Bandpass of Preamplifier (40 dB) with Vacuum Grease Couplant

Sensor / Wave Form Feature	Average for First Half Cycle of Extensional Mode*, dB	Average for First Half Cycle of Flexural Mode+, dB
Harisonics, Model G-0504 1/4 inch diameter crystal	-12	-16
Valpey Fisher, Pinducer Model 1093, 0.054 inch diameter crystal	-25	-23
EBL, Model NBS Point Contact, ~0.025 inch	-15	-19
Physical Acoustics, Model μ-30 Crystal Diameter ~0.31 inch	0 (reference)	0 (reference)

* 0.3 mm lead break at center of plate edge

\+ 0.3 mm lead break at top of plate edge

Table 4. Approximate Effect of Digitization Rate on Accuracy of Determined Group Velocities, Range and Direction *

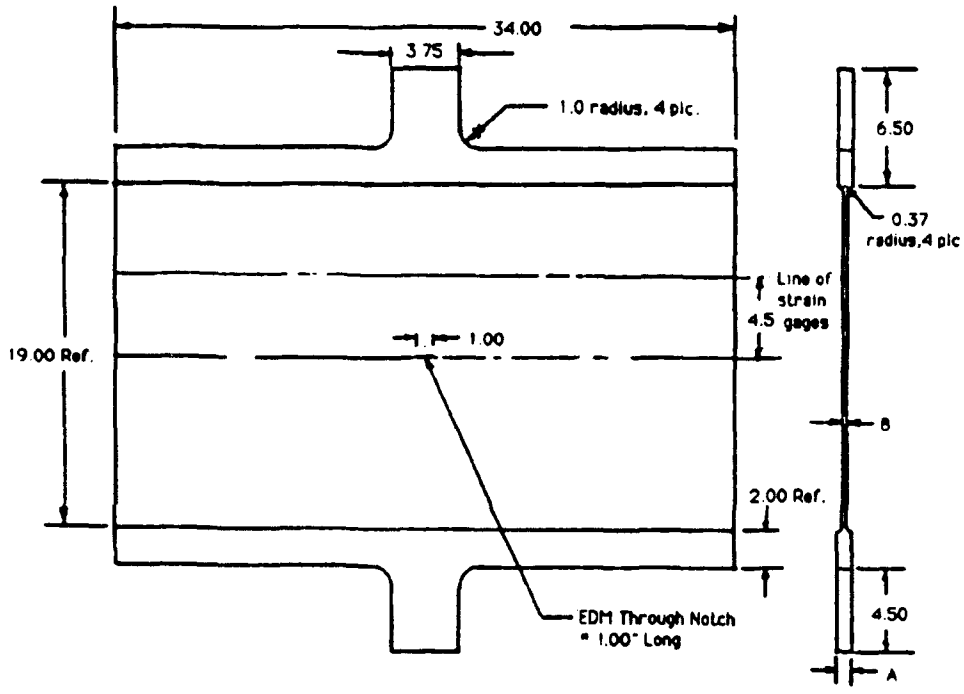
Digitization Interval, μs	Percentage Error			
	Higher Group Velocity	Lower Group Velocity	Range to Source	Direction to Source
0.5	0.5	2.2	3.4	7.4
0.3	3.4	2.6	11.3	6.7
0.1	0.6	1.2	1.7	0.2
0.05	0.5	0.3	1.1	0.2
0.03	0.5	0.6	0.7	0.01
0.010	0.1	0.2	0.2	0.06

* Assumed higher group velocity 0.214 in/μs, lower group velocity 0.114 in/μs, range to center of SAT 6.0 in, and angle with respect to SAT axis 30°.

Table 5. Loading Conditions for AE Test Platforms

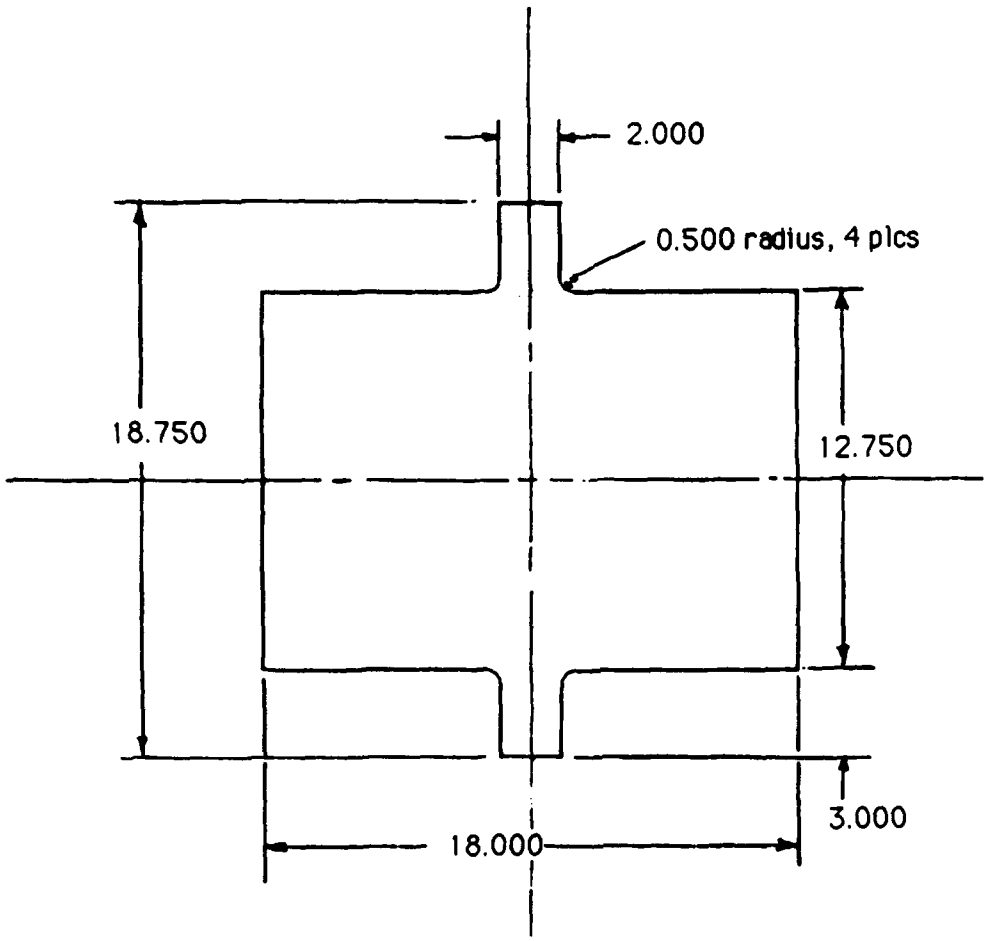
Plate Section Thickness, in	Interrupted Cycling		Overload Test	Higher Rate Cycling	
	Maximum * Load,lbs	Crosshead Rate, in/min.	Maximum Load,lbs	Maximum * Load,lbs	Crosshead Rate, in/min.
1/8	9,100	0.002	10,000	10,500	0.2
1/4	16,900	0.002	30,000	--	--

* Minimum < 1% of maximum



Dim./Tab	1	2
A	0.750	0.625
B	0.250	0.125

Figure 1. Drawing of nominal dimensions of solid AE test platform (All dimensions in inches).



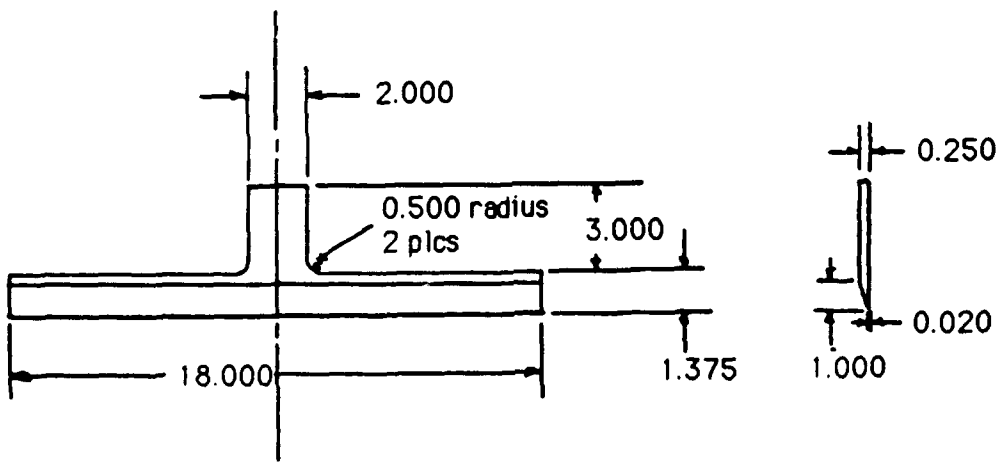


Figure 2. Drawing of nominal dimensions of sub-scale adhesively bonded sample showing main plate and reinforcement tabs to be bonded to the ends of the sample (All dimensions in inches).

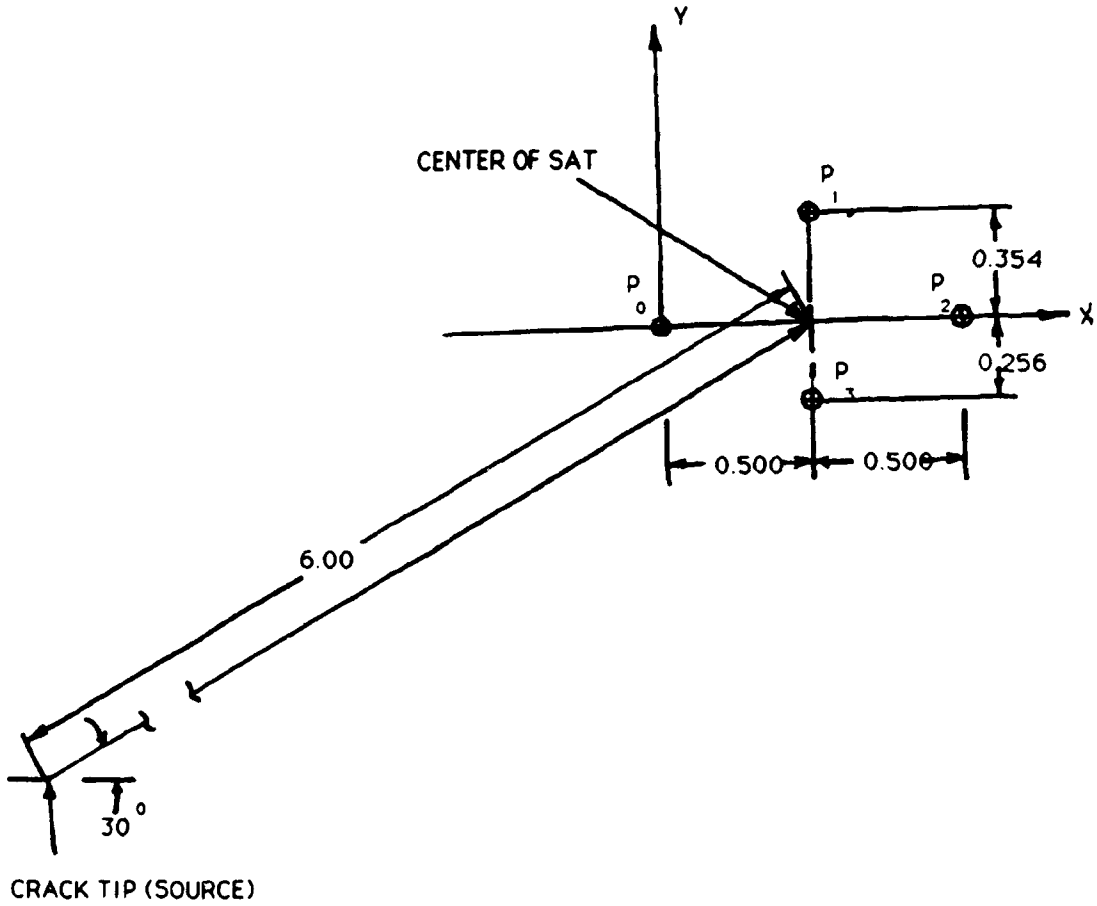


Figure 3. Geometry of SAT relative to simulated and crack tip source (All dimensions in inches, $p_0 - p_3$: pinducer positions).

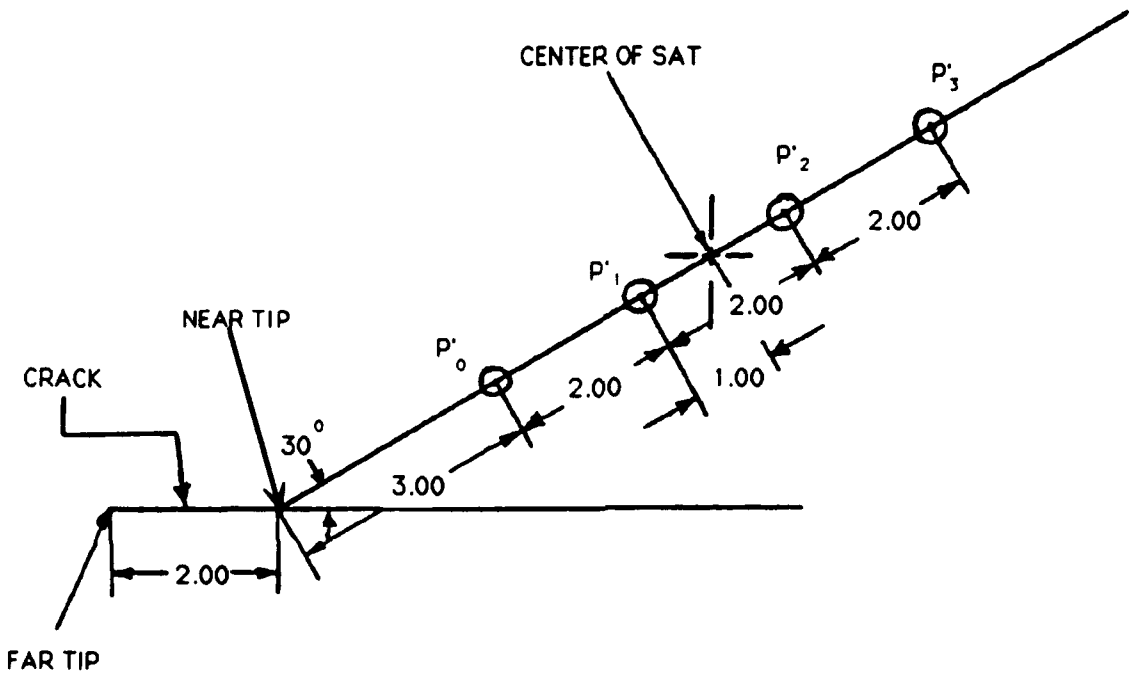


Figure 4. Relative positions of SAT and in-line pinducer configurations (All dimensions in inches, $p'_0 - p'_3$: pinducers).

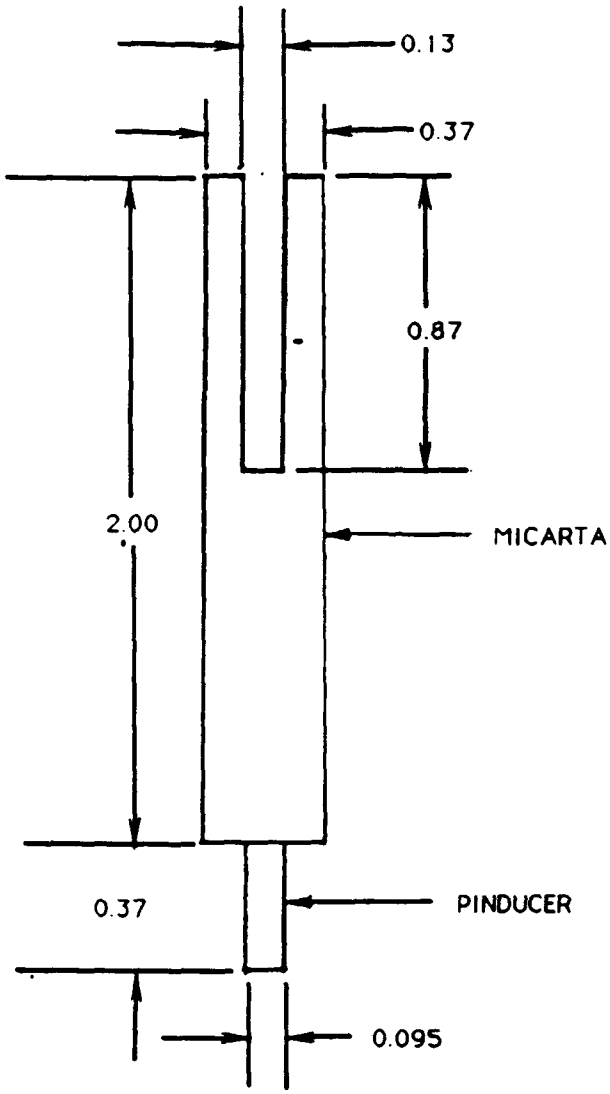


Figure 5. Drawing of pinducer without connector adhesively bonded into micarta holder (All dimensions in inches).

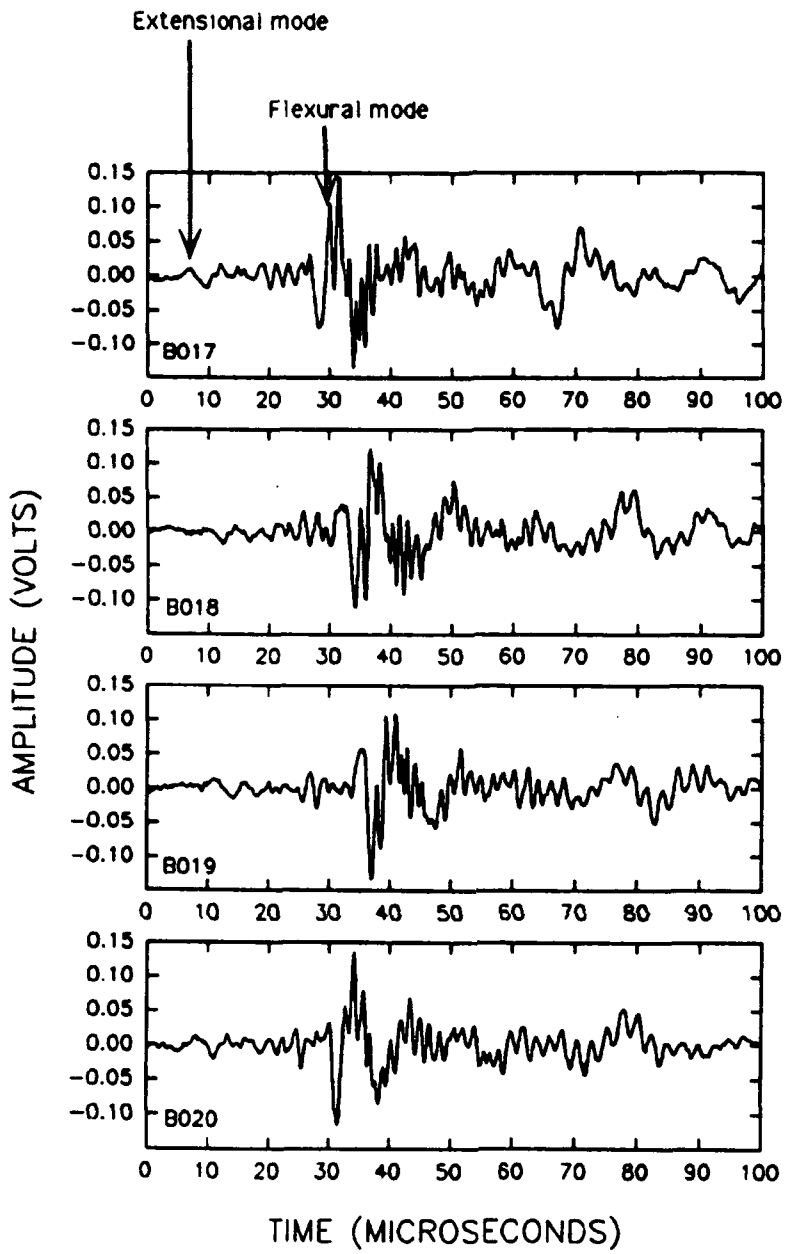


Figure 6 (a). Large signal to noise waveforms from SAT configuration for crack opening source in center-crack plate (1/4 inch thick plate).

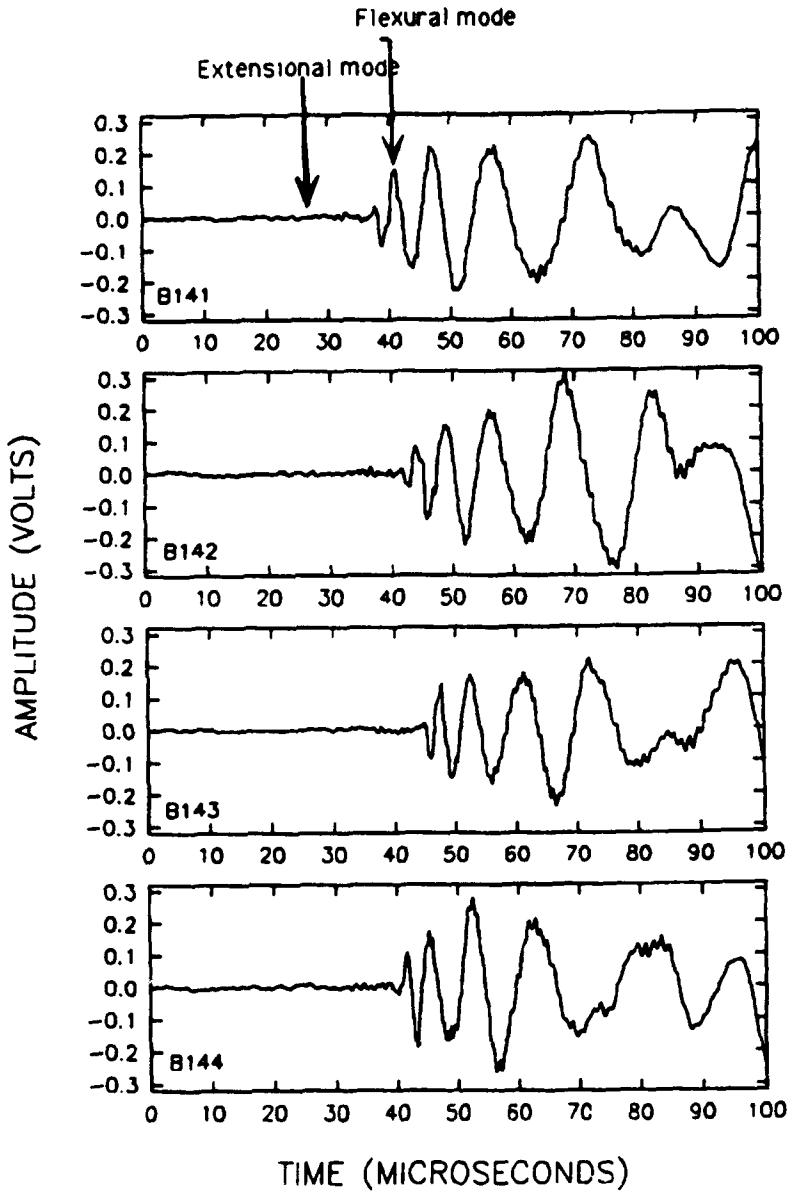


Figure 6 (b). Large signal to noise waveforms from SAT configuration for crack opening source in center-crack plate (1/8 inch thick plate).

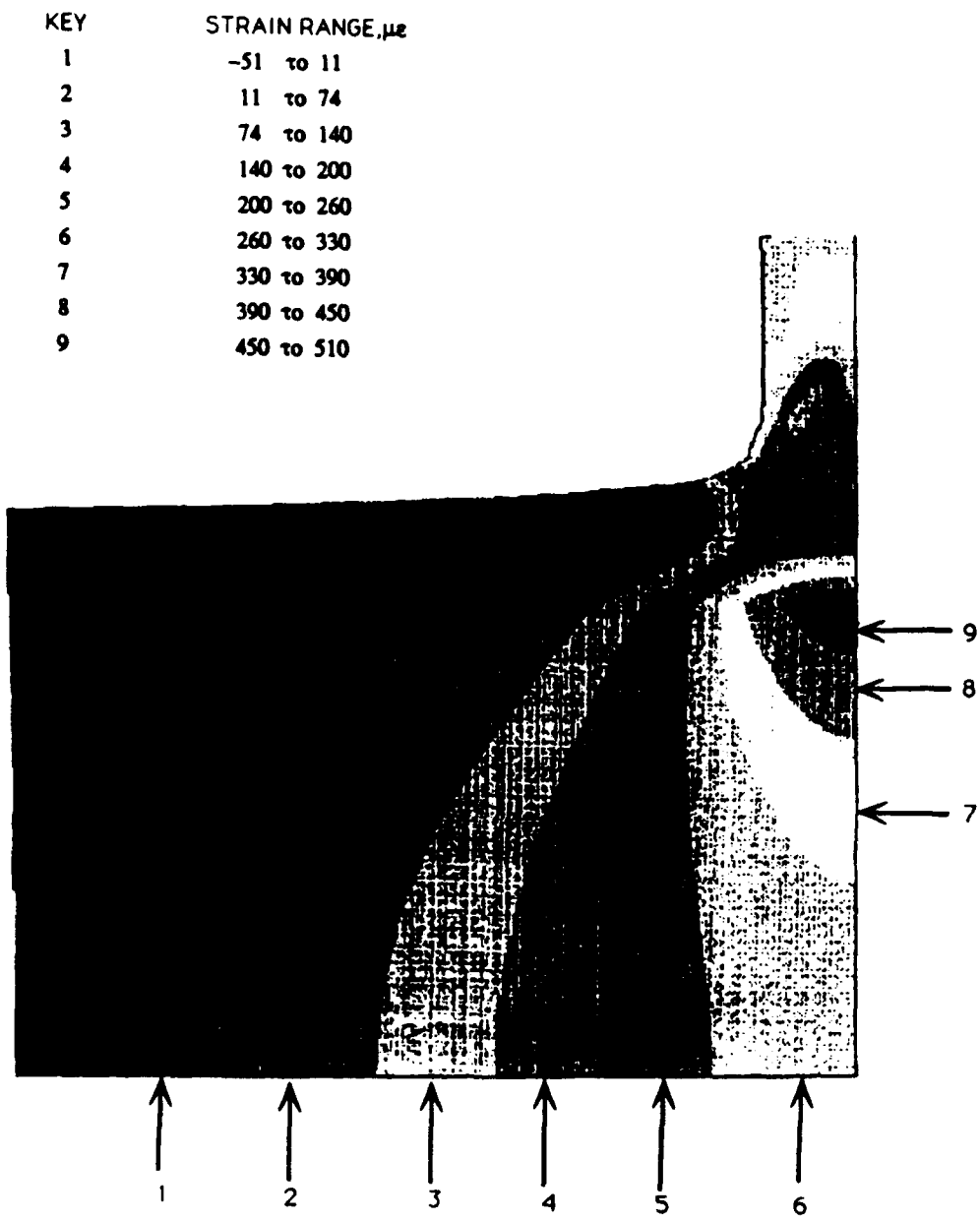


Figure 7. Finite element axial strain distribution in one quadrant at a load of 10,000 lbs for 1/4 inch thick test platform without a crack.

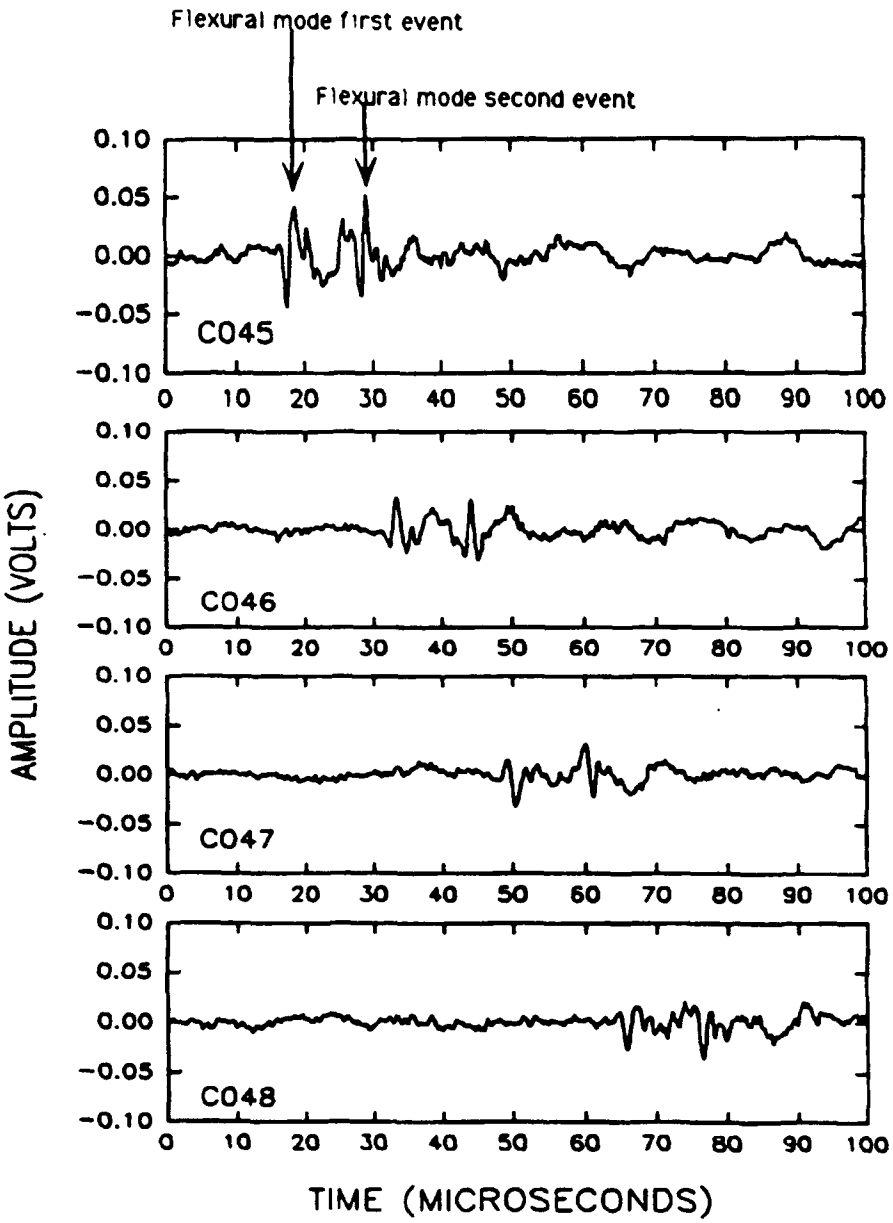


Figure 8. Two events separated by 10.8 μs for crack opening source in 1/8 inch plate with in-line sensor configuration at 2430 lbs.

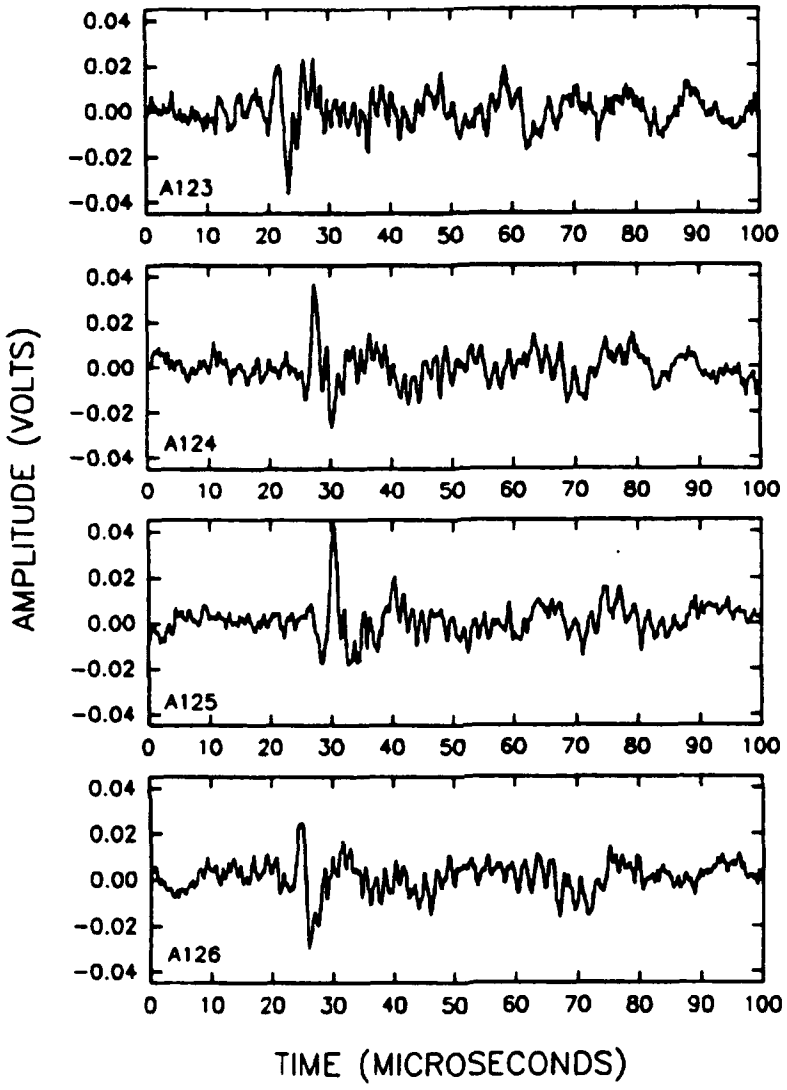


Figure 9. Waveforms from SAT with typical signal to noise ratio from crack opening event at 2430 lbs load on the 1/4 thick plate.

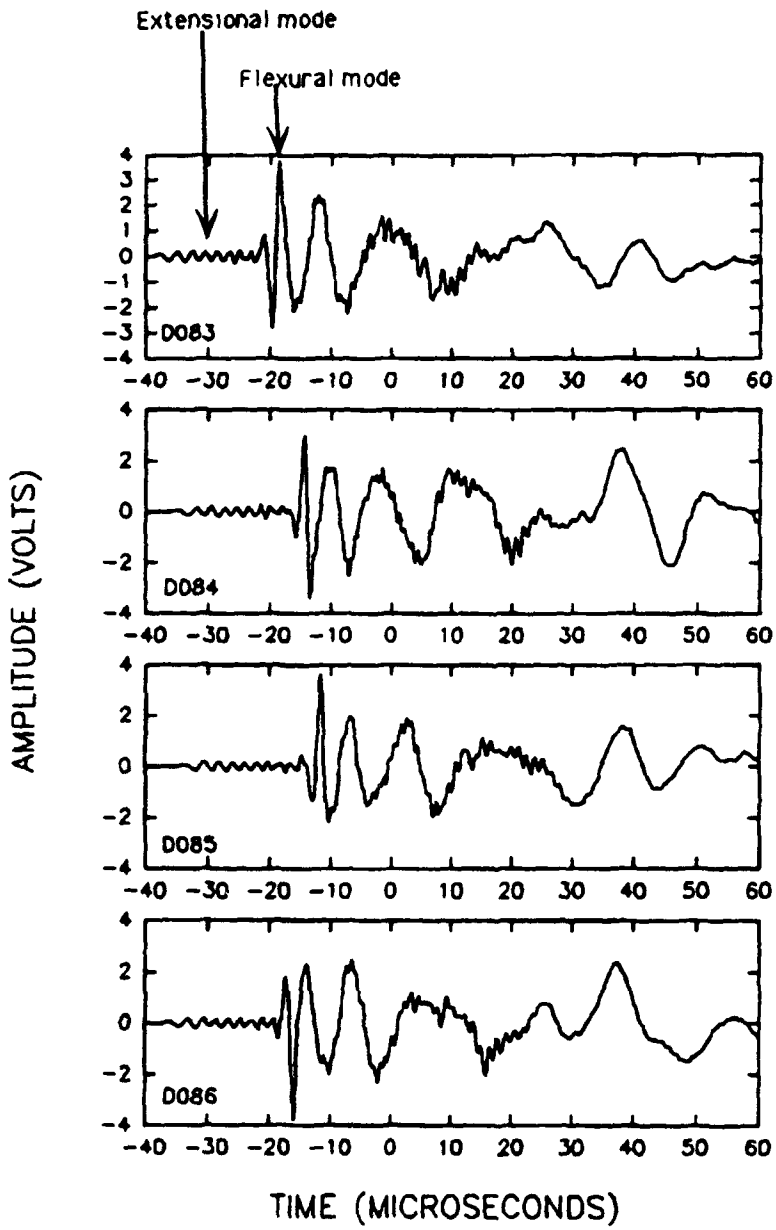


Figure 10 (a). Waveforms from SAT for lead break at near crack tip on 1/8 inch thick plate (Gain 40 dB).

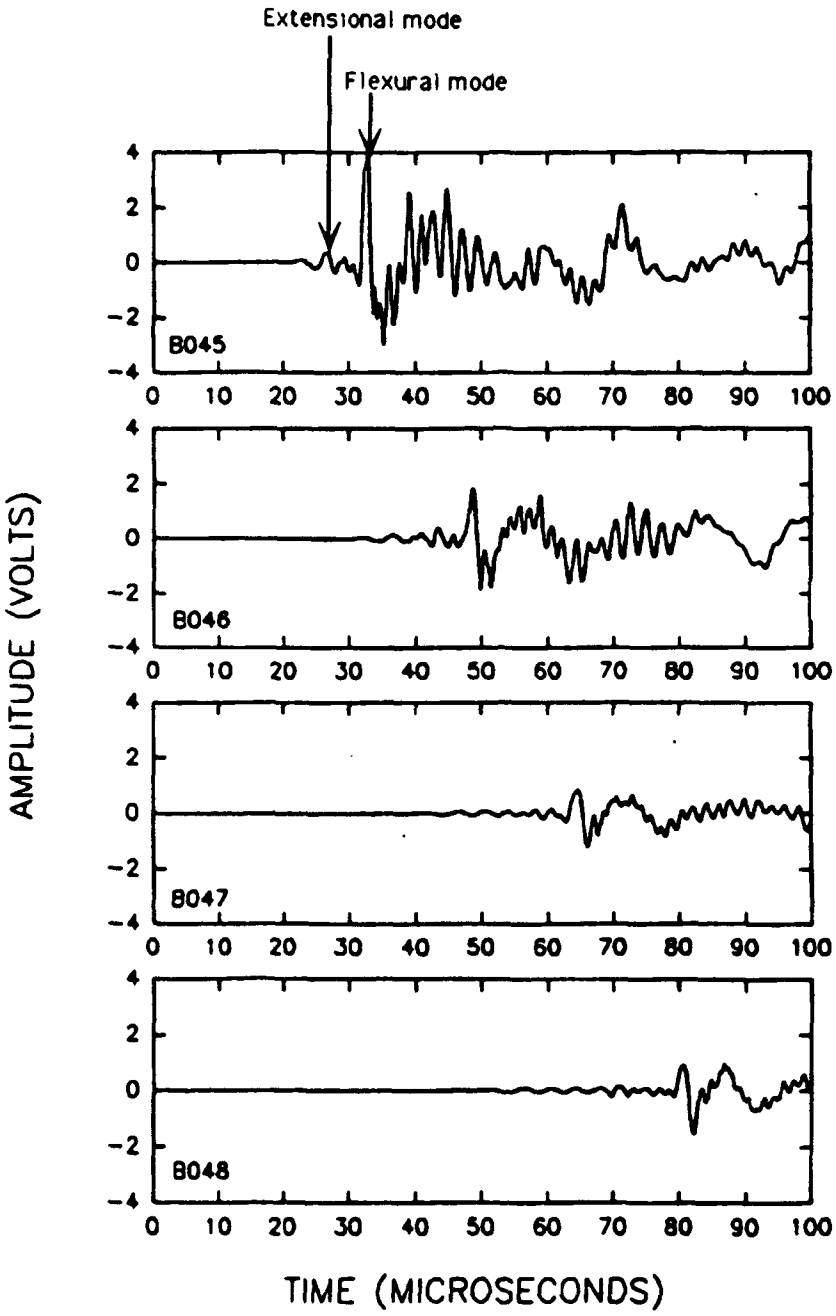


Figure 10 (b). Waveforms from in-line configuration for lead break at near crack tip on 1/4 inch thick plate (Gain 40 dB).

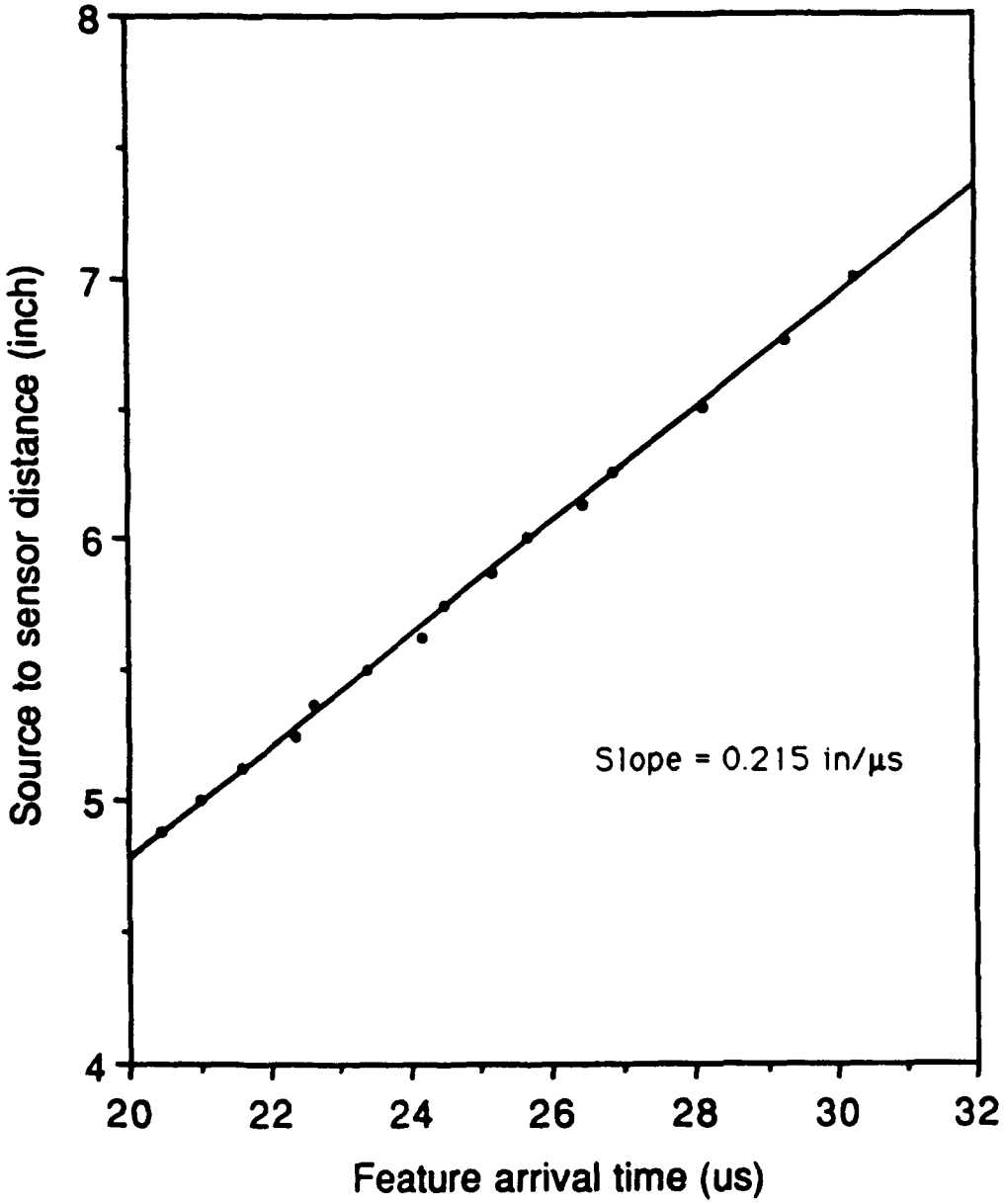


Figure 11. Source distances versus first half cycle peak extensional feature arrival time for top edge lead breaks on 1/8 inch thickness sample.

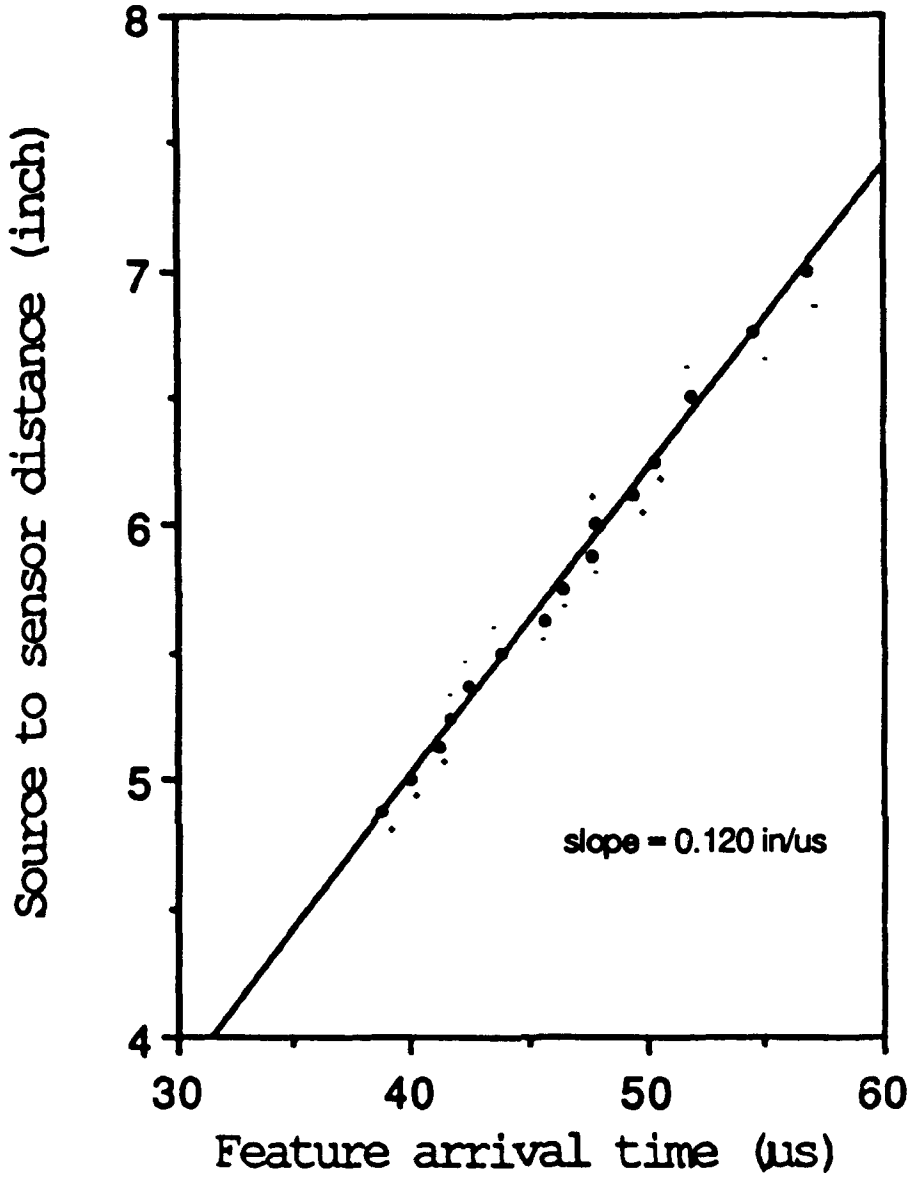


Figure 12. Source distance versus arrival time of zero crossing of high amplitude flexural feature with the steepest slope on 1/8 inch thickness sample (sign indicates slope of zero crossing).

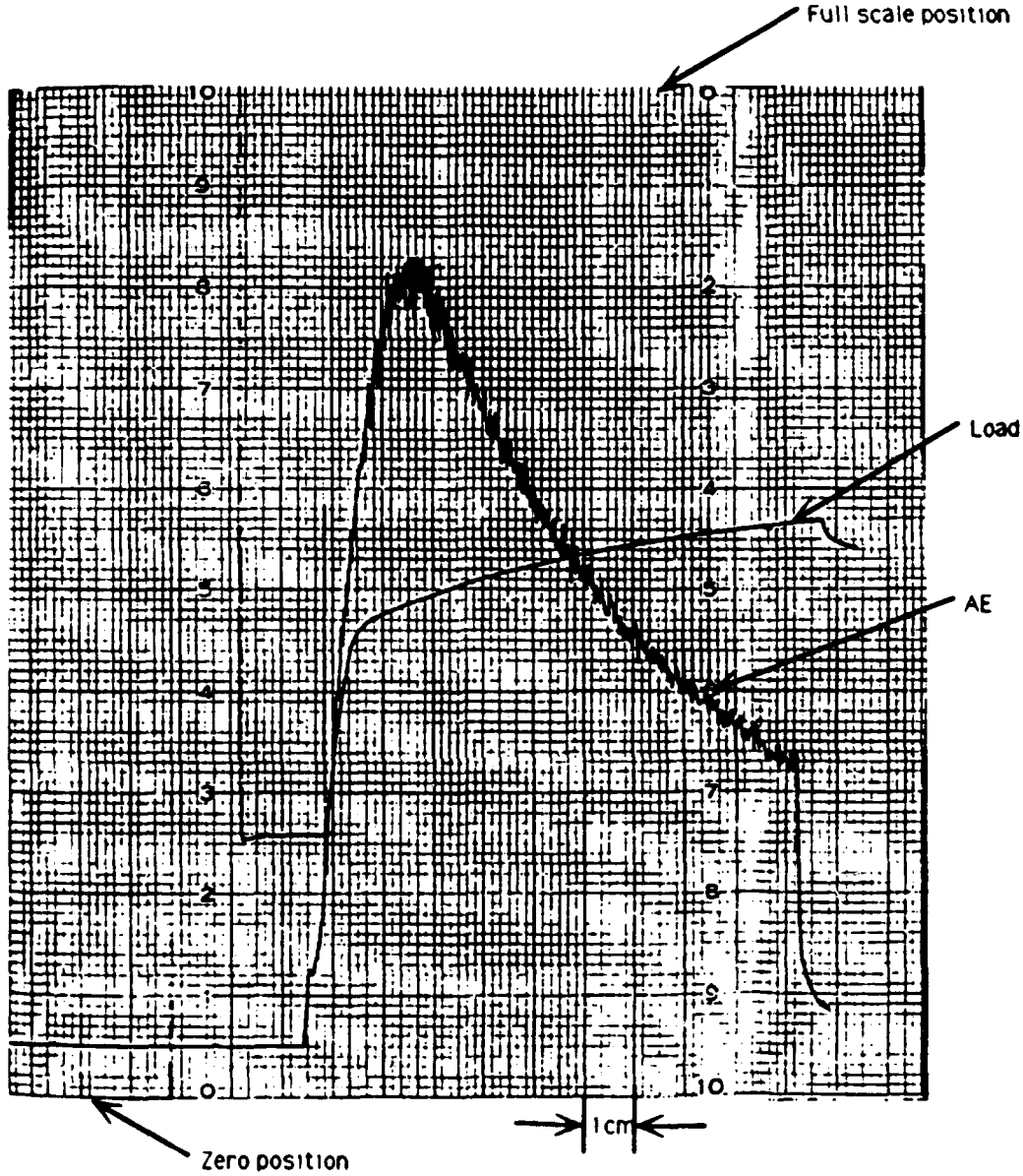


Figure 13 (a). Root-mean-square of AE (88 dB & 100-300 kHz) and load versus time for tensile test of 2024-T351 used in reference 2 (Full scale values are RMS = 0.1V; Load = 5,000 lbs; and chart speed = 4 cm/min).

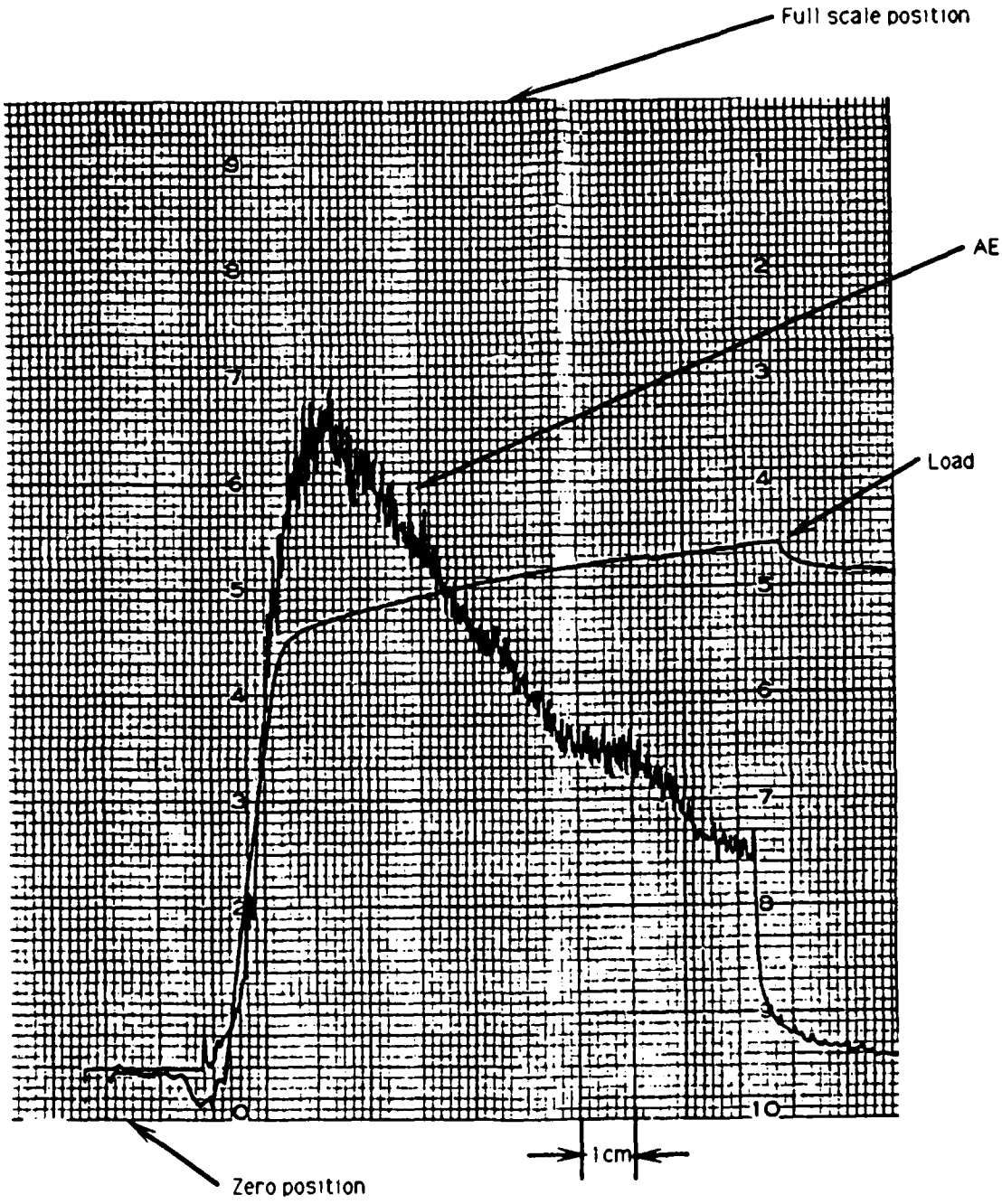


Figure 13 (b). Root-mean-square of AE (88 dB & 100-300 kHz) and load versus time for tensile test of 2024-T351 from 1/4 inch thick plate (Full scale values are RMS = 0.1V; Load = 5,000 lbs; and chart speed = 4 cm/min).

APPENDIX I. Short Report from Dr. Gorman

Location of Acoustic Emissions From Cold Weld-Type Sources in Fatigue Cracks in 2024 Aluminum

Michael R. Gorman
Department of Aeronautics and Astronautics
Naval Postgraduate School
Monterey, California 93943
(408) 646-2074

Preliminary Report
27 December 1991

Introduction

This report is an added appendix to the main report by Dr. Marvin A. Hamstad. The purpose of this section is to give the results of a different method of locating and analyzing acoustic emission caused by fracturing of tiny cold welds behind crack tips in aluminum plates. The details of the plates, the creation of fatigue cracks, and the loading schemes are given in the main report.

Experimental

Three broadband transducers (Harisonic, Model G0405) were affixed with Tacki-wax (Cenco) to the surface of the plate in a triangular arrangement about the crack. It was desired to measure the vertical displacement of the surface as the transient pulse passed underneath the transducer. The transducers were surface contact ultrasonic types having a broad resonance at 5 Mhz. They were selected because they have been shown to have a smooth and nearly flat response below one megahertz where acoustic emission is found. The signals from the transducers were fed into wideband preamplifiers which, in turn, were connected to transient recorders. The digitized waveforms were stored on a hard disk in a personal computer.

Typically, narrowband transducers have been used for acoustic emission work because of their high sensitivity. Source location is carried out by starting and stopping AE analyzer clocks when the signals at the transducers first cross a preset threshold. It has been shown in earlier work by this author that such a method can lead to serious errors in source location [1]. A different source location technique

has been developed called Gaussian crosscorrelation which takes the wave propagation in a plate into account and, in principle, is capable of high accuracy [2]. It is not based on threshold crossing. It does require high fidelity waveform capture however and thus the use of broadband transducers. At kind invitation of Dr. Marvin Hamstad and Dr. George Sendeckyj I was allowed to piggyback on top of their experiment to see if this new technique would work with real AE sources (as opposed to artificial sources like lead breaks).

Results

The AE pulses created by the cold weld fracture were small. The flexural mode was usually all that was detected. This is a highly dispersive mode in a thin plate which means that it changes shape rapidly as it propagates. Generally only a few events were detected in each loading. Representative results using the flexural mode are given in the figures for the two different plates (1/4" and 1/8").

Figure 1 shows the measured locations of two 0.3 mm lead breaks done at the far crack tip and one done at the near tip on the 1/4" plate. The actual break locations were as close to the tips as could be determined by eye. 80 Khz was selected for the location calculations since this appeared in all of the waves which could be analyzed. Other frequencies can also be used. The box lines represent the plate and the crack is shown in scale.

Figure 2 shows the measured locations of actual AE events. It can be seen that the transducers were placed in a different arrangement than in Test 1. Different arrangements of the transducers were tried for various reasons having mainly to do

with desiring to see the shape of the pulse at different angles to the crack plane. No effect on source location was expected and none was found but it probably wouldn't be noticed with the actual crack events since the source position was known only to be somewhere behind the crack tips.

Figure 3 shows the waveforms from one event as measured by the three sensors shown in Figure 2. This was the best looking set of waveforms of the five measured. Figure 4 gives an idea of a more typical set. It was difficult to extract the information if they were smaller than this.

A result for the 1/8 plate is shown in Figure 5. There were only two events measured during this particular loading.

Conclusions

The results thus far are quite encouraging. The signals were very small and contained a lot of noise but locations could be obtained for some of the larger events. The accuracy cannot be determined because the exact location of any particular source itself was not known (it was only known that the source was in the crack plane somewhere behind the crack tips). The events were all placed near the crack to within about one-half of an inch. This is within the error created by the method used to position the sensors and the quarter-inch size of the sensors.

References

1. Gorman, M. R. and Ziola, S. M., "Plate Waves Produced by Transverse Matrix Cracking," *Ultrasonics*, vol. 29, p. 245 1991.

2. Ziola, S. M. and Gorman, M. R., "Source Location in Thin Plates Using Cross-correlation," *Journal of the Acoustical Society of America*, vol.29 (5), 1991.

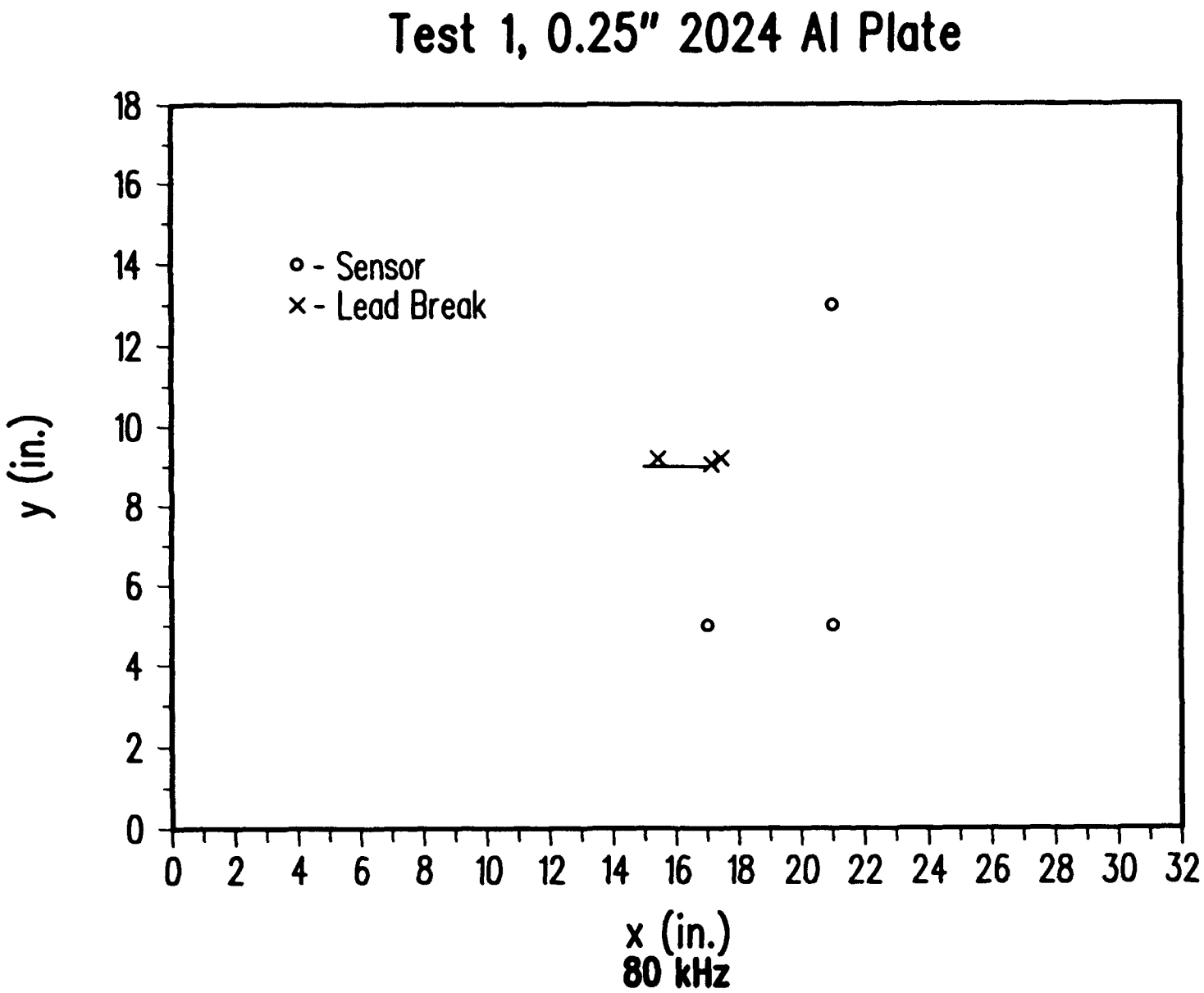

Test 1, 0.25" 2024 Al Plate
o - Sensor
x - Lead Break
y (in.)
x (in.)
80 kHz
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32

Figure 1.

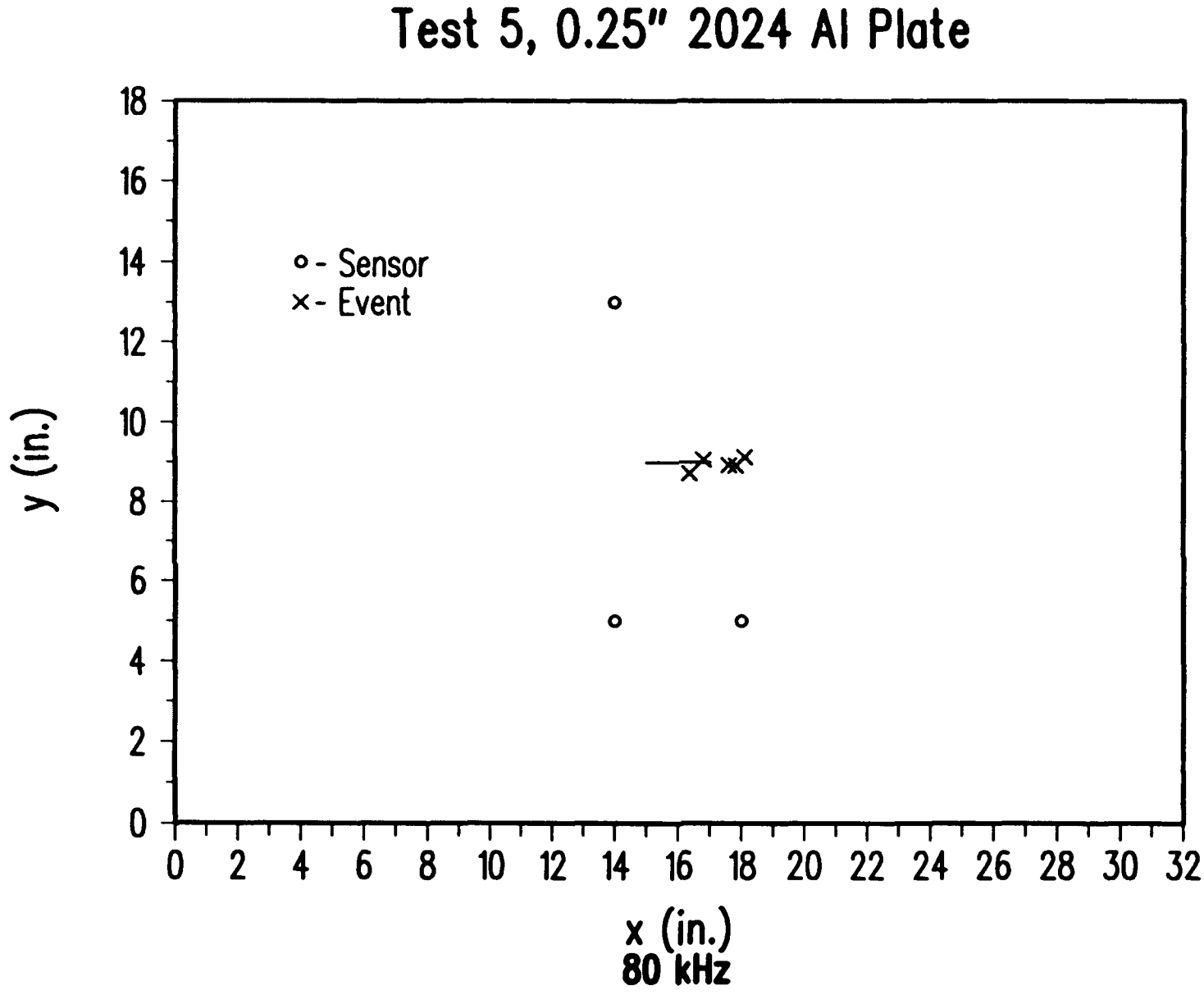
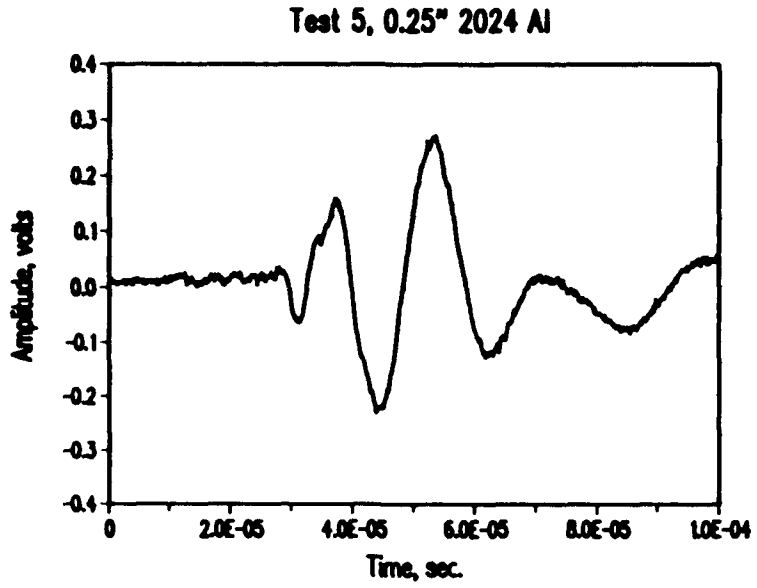
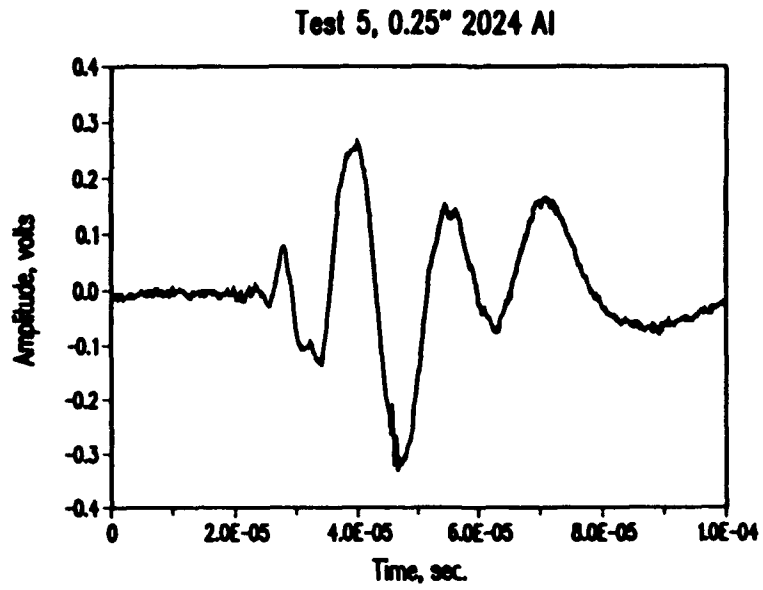

Test 5, 0.25" 2024 Al Plate
o - Sensor
x - Event
y (in.)
x (in.)
80 kHz
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32

Figure 2.


Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.


Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.

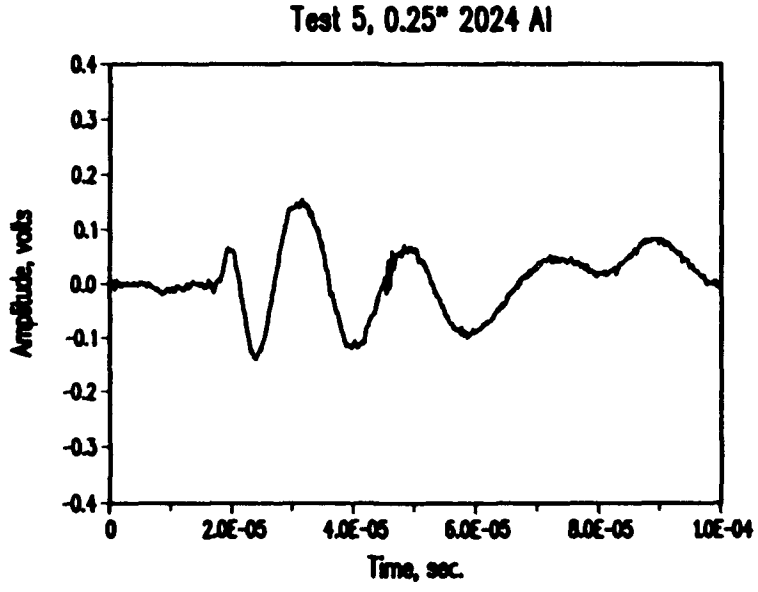
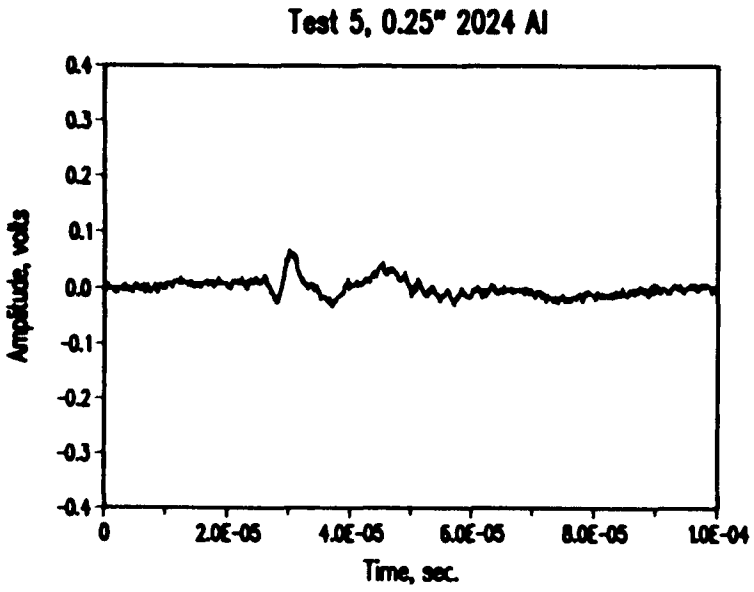
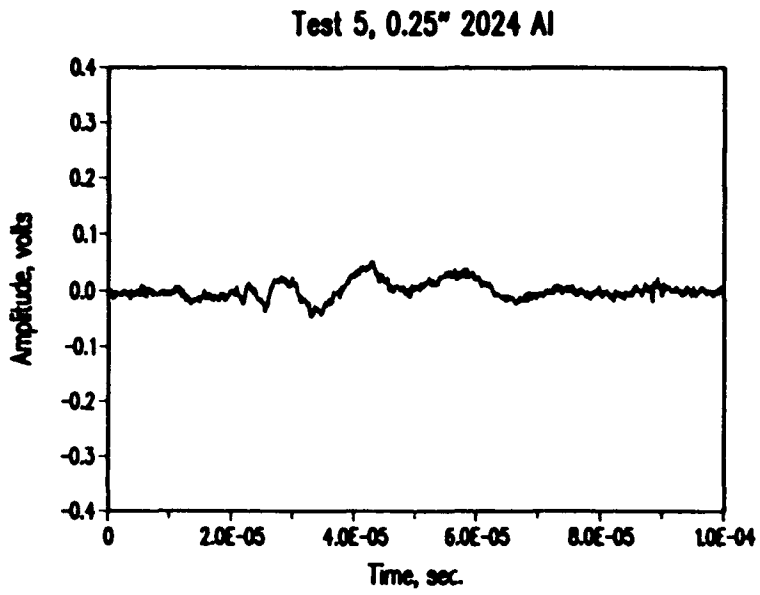

Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.

Figure 3.


Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.


Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.

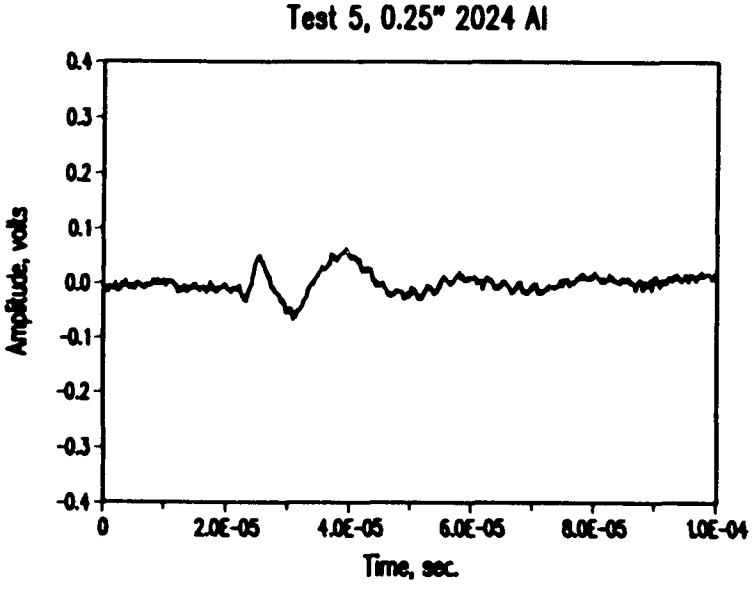

Test 5, 0.25" 2024 Al
Amplitude, volts
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
2.0E-05
4.0E-05
6.0E-05
8.0E-05
1.0E-04
Time, sec.

Figure 4.

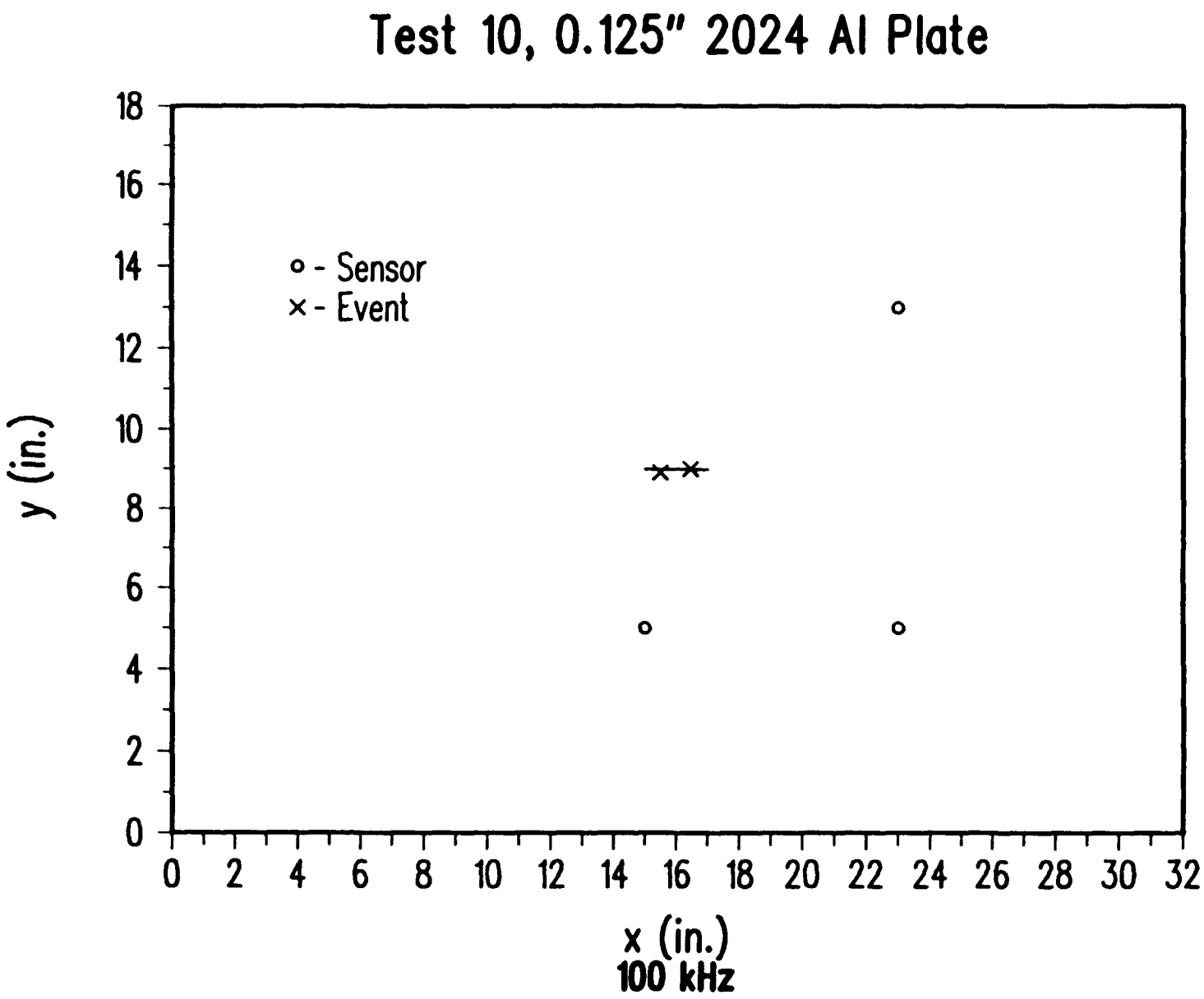

Test 10, 0.125" 2024 Al Plate
o - Sensor
x - Event
y (in.)
x (in.)
100 kHz
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32

Figure 5.

APPENDIX II. Equations and their Partial Modification from the SAT Patent

From equation (31) in the patent for each wave feature

$$a=\frac{T_2-T_0}{x_2},\ b=\frac{T_1-T_3}{y_1+|y_3|} \qquad \text{II-1}$$

where from figure 4 in the array coordinate system $x_1 = x_3 = 0.500$ in, $x_o = o$, $x_2 = 1.000$ in., $y_o = y_2 = o$, $y_1 = 0.354$ in., $y_3 = -0.256$ in. and T_o, T_1, T_2, T_3 are the relative arrival times of the selected feature at sensors p_o, p_1, p_2, p_3, respectively.

From equation (36) in the patent corrected for a missing minus sign in the denominator

$$R_o=\frac{\Delta T}{(a_2^2+b_2^2)^{1/2}-(a_1^2+b_1^2)^{1/2}} \qquad \text{II-2}$$

where R_o is the range from the source to a sensor (e.g. to sensor p_o) and ΔT is the arrival time difference between two selected features with subscripts 1 and 2. The a and b values are calculated from equation II-1.

From equation (32) in the patent the group velocities of the two selected features are given by

$$c=\frac{1}{(a^2+b^2)^{1/2}} \qquad \text{II-3}$$

where the appropriate subscript is attached.

Since in the current work the typical range from the source to the sensors is not much greater than the spacing of the sensors, the simplification used in the patent to go from equation (24) to equation (30) leads to errors in the determination of the components of the unit vector $\vec{u}_R$: $[u_x, u_y, o]$ from the source to the sensor p_o (Note: Following the patent, we ignore plate thickness and determine the two-dimensional source position in this work.) To obtain accurate results, equation (30) from the patent was used to find trial values for u_x and u_y (using both of the features and averaging the results).

$$u_x=\frac{a}{(a^2+b^2)^{1/2}} \qquad \text{II-4(a)}$$

$$u_y=\frac{b}{(a^2+b^2)^{1/2}} \qquad \text{II-4(b)}$$

Then from equation (29) in the patent (corrected)

$$\vec{u}_R \cdot \vec{u}_i = \frac{\delta_i}{|\vec{r}_i|} = [1/2(1+\frac{R_i}{R_o})] - [\frac{r_i}{2R_o}] \qquad \text{II-5}$$

Where $\vec{u}_i$ is the unit vector in the sensor coordinate system (with origin at p_o) to each sensor p_1, p_2, p_3 for i = 1, 2, 3, respectively, $R_i = |\vec{R}_i|$ is the range from the source to each sensor, $r_i = |\vec{r}_i|$ is the range from sensor p_o to p_1, p_2, p_3, for i = 1, 2, 3, respectively, and δ_i is given by equation (9) from the patent

$$\delta_i = (T_i - T_o)c \qquad \text{II-6}$$

where c is given by equation II-3 for a selected feature. Also from the patent equation (1).

$$\vec{R}_i = \vec{R}_o + \vec{r}_i \qquad \text{II-7}$$

Now after some algebraic manipulations, equations II-5 can be solved for

$$u_x = \frac{1}{x_2}\{\delta_2[1/2(1+\frac{R_2}{R_o})] - \frac{r_2^2}{2R_o}\} \qquad \text{II-8(a)}$$

$$u_y = \frac{1}{y_1 - y_3}\{\delta_1[1/2(1+\frac{R_1}{R_o})] - \delta_3[1/2(1+\frac{R_3}{R_0})] - \frac{r_1^2}{2R_o} + \frac{r_3^2}{2R_o}\} \qquad \text{II-8(b)}$$

Now since $\vec{r}_1:[x_1,y_1]$, $\vec{r}_2:[x_2,y_2]$, $\vec{r}_3:[x_3,y_3]$ then using the trial values of u_x and u_y from equation II-4 and R_o from equation II-2, the values of R_i can be found from equation (2) in the patent, i.e.

$$R_i^2 = (\vec{R}_o + \vec{r}_i)\cdot(\vec{R}_o + \vec{r}_i) \qquad \text{II-9}$$

Then equations II-8 yield two new values for u_x and u_y (one for each feature) with group velocities c_1 and c_2. These values were averaged and taken as new trial values, and the process was repeated one more time, resulting in convergence to the correct value for u_x and u_y.

1990 USAF-UES Research Initiation Program

Sponsored by
Air Force Office of Scientific Research
Conducted by
Universal Energy Systems, Inc.

Final Report
H-infinity Control Design--A New Approach

Prepared by: Chin S. Hsu, Ph.D. Chin HS

Academic Rank: Associate Professor

Department: School of Electrical Engineering and Computer Science

University: Washington State University
Pullman, WA 99164-2752

USAF: Flight Dynamics Laboratory
WRDC/FIGC
Wright-Patterson AFB, OH 45433

Date: December 27, 1991

Contract No: F49620-88-C-0053

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. Universal Energy Systems must be mentioned for their concern and help in all administrative and directional aspects of this program.

My experience on this research was rewarding and enriching. I would like to thank Mr. John Bowlus, Dr. Siva Banda and Dr. Hsi-Han Yeh for their technical guidance, support and keen interest in my research. I also wish to thank Dr. Jenny Rawson and Mr. Xianggang Yu for their technical assistance in various aspects of my research efforts.

H-infinity Control Design--A New Approach

by
Chin S. Hsu
Washington State University

Abstract

This report addresses the issue of the H∞ compensator design based on full order and reduced order observers. A new approach for robust control design is proposed. New results pertaining to approximate and precise H∞ loop transfer recovery (H∞/LTR) have been derived. Numerical examples illustrating the proposed design procedures are also presented.

H-infinity Control Design--A New Approach

I. Introduction

Recent progress on H_∞ control theory has provided control engineers with practical design procedures for developing robust control laws. Though different approaches of H_∞ suboptimal design have been devised, the resultant compensators are, in general, observer-based.[1] These H_∞ design approaches include the celebrated two-Riccati formulation, the conjugation method, the bounded-real-lemma method, J-spectral factorization method among others.[2] Along with the development of H_∞ output feedback design, H_∞ state feedback and its dual, H_∞ filtering, have attracted considerable attention during past few years.[3]

Motivated by the recent results of disturbance rejection, we propose a new method of deriving an H_∞ compensator design.[4] The new method is based on the notion that all H_∞ suboptimal controllers are observer-based. One significant benefit of the proposed approach is that it provides reduced-order H_∞ controllers based on reduced-order observers.

Considerable research has been carried out by researchers for the control law synthesis using the LQG/LTR and the H_2/H_∞ methodologies.[5] The challenging problem of H_∞/LTR has also been investigated in recent years.[6],[7] Influenced by the foregoing robust control design methods, we developed control design procedures and numerical algorithms which provide H_∞ compensators based on full-order or reduced-order observers. Several results pertaining to approximate and precise H_∞/LTR are presented in this report.

To facilitate the readability of this report, we defer the mathematical derivation of the main results to the Appendix at the end of this report. Notations and block diagrams of various control sytems configurations are given

in Section III for the ease of relating the theoretical results with the physical setting. The key results of this research are collectively presented in Section II, while numerical examples illustrating the key results can be found in Section IV.

Finally, Section V offers some concluding remarks and recommendations for future research.

II. Summary of Main Results

2.1 Full State Feedback

We begin with the control system configuration (Figure 1) which is used for designing H_∞ suboptimal controllers, where S and S_o represent the plant and its associated observer, respectively. The H_∞ design is concerned with developing a regulator K and an observer S_o such that H_∞ norm of the transfer function matrix from w to z is smaller than a prespecified bound, while the closed-loop system is internally stable.

The plant S is represented by the following linear system:

$$\begin{aligned} \dot{x}(t) &= A\,x(t) + B_1\,w(t) + B_2\,u(t), \quad x(0) = x_o \\ z(t) &= C_1\,x(t) + D_{11}\,w(t) + D_{12}\,u(t) \\ y(t) &= C_2\,x(t) + D_{21}\,w(t) \end{aligned} \tag{1}$$

where z(t) is a vector of performance variables and w(t) represents a deterministic disturbance. The dimensions of x(t), u(t), z(t), and y(t) are n, m. ℓ, and p, respectively. The following assumptions are standard in H_∞ design.

$\{A, B_1\}$, $\{A, B_2\}$ stabilizable

$\{A, C_2\}$, $\{A, C_1\}$ detectable

$C_1^T D_{12} = 0$, $D_{12}^T D_{12} > 0$, $B_1 D_{12}^T = 0$

Lemma1

The closed-loop transfer function matrix (TFM) from w to z is

$$T_{sF}(s) = D_{11} + (C_1 - D_{12}K)(sI - A + B_2K)^{-1} B_1 \tag{2}$$

if state feedback is used, i.e.,

$$y(t) = x(t) \text{ and } u(t) = -K\,x(t)$$

It is known that a gain matrix K can be obtained by solving a modified Riccati equation, to nearly minimize the H_∞ norm of $T_{SF}(s)$.[8] Suppose that $X_\infty > 0$ is a solution to

$$0 = X_\infty A + A^T X_\infty + X_\infty(\gamma^{-2} B_1 B_1^T - B_2(D_{12}^T D_{12})^{-1} B_2^T) X_\infty + C_1^T C_1 \tag{3}$$

such that $A - B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty$ is stable. Then, if the regulator gain is computed as,

$$K = (D_{12}^T D_{12})^{-1} B_2^T X_\infty \tag{4}$$

the norm of the closed-loop transfer function matrix is bounded by γ:

$$\|T_{SF}(s)\|_\infty < \gamma. \tag{5}$$

2.2 Full-Order-Observer Based Controller

A full-order observer S_o is of the form,

$$\dot{x}_c(t) = (A - LC_2)\,x_c(t) + Ly(t) + B_2u(t) \tag{6}$$

$$u(t) = -K\,x_c(t), \qquad x_c(t) \in R^n$$

where L denotes the observer gain.

Theorem 1 [9]

The observer-based closed-loop TFM from w to z is

$$T(s) = T_{SF}(s) + F(s)\,(sI - A + LC_2)^{-1}\,(B_1 - LD_{21}) \tag{7}$$

where F(s) is a stable filter described by

$$F(s) = D_{12}K + (C_1 - D_{12}K)\,(sI - A + B_2K)^{-1}\,B_2\,K \tag{8}$$

By examining the result of Theorem 1, we see that once K is obtained by solving an H_∞ state feedback problem, F(s) is a known TFM of a linear system. In fact, equation (7) is of the form as the well known one-sided model matching problem (MMP), if (7) is rewritten as

$$T(s) = T_{SF}(s) + F(s)M(s) \tag{9}$$

where

$$M(s) \triangleq (sI - A + LC_2)^{-1} (B_1 - LD_{21}) \tag{10}$$

It is noted here that in (9), both transfer function matrices $T_{SF}(s)$ and F(s) are stable. The link between H_∞ compensation design and stable model matching problem is an important one. As to be presented in the sequel, this connection plays a vital role in developing precise H_∞/LTR design procedures. Moreover, the stable filter F(s) can be considered as a frequency-shaped filter which predicates the determination of the observer gain matrix L, (Figure 2).

The unknown parameter L of (7) is of a form dual to its counterpart K of (2). Hence, an H_∞ output feedback compensator can be arrived at by determining the observer gain L such that

$$\begin{aligned} &\| T(s) - T_{SF}(s) \|_\infty \\ &= \| F(s) (sI - A + L C_2)^{-1} (B_1 - LD_{21}) \|_\infty \\ &= \| F(s)M(s) \|_\infty < \delta \end{aligned} \tag{11}$$

where δ is a positive number.

It is found that the selection of L is essentially a frequency-shaped state estimator problem. This is solved according to the following theorem.

<u>Theorem 2</u>: Let $Z_\infty > 0$ be a solution to

$$AZ_\infty + Z_\infty A^T + Z_\infty(\delta^{-2}X_\infty B_2(D_{12}^T D_{12})^{-1}B_2^T X_\infty - C_2^T(D_{21}D_{21}^T)^{-1}C_2)Z_\infty + B_1 B_1^T = 0 \tag{12}$$

such that $A - Z_\infty C_2^T(D_{21}D_{21}^T)^{-1}C_2$ is stable and $\delta > 0$. Then,

$$L = Z_\infty C_2^T(D_{21}D_{21}^T)^{-1} \tag{13}$$

gives $\|T(s) - T_{SF}(s)\|_\infty < \delta$.

It is seen from equation (13) that the resulting closed-loop transfer function matrix T(s) is approximately equal to the closed-loop transfer function matrix with H_∞ state feedback. To summarize the main results presented so far, we give a two-step design procedure:

1) Determine a state feedback gain K such that $\|T_{SF}(s)\|_\infty < \gamma$.

2) Determine an observer gain L such that $\|T(s) - T_{SF}(s)\|_\infty < \delta$.

Each step of the design procedures requires the solution of only one Riccati equation.

2.3 Reduced-Order-Observer Based Controller

Let us consider again the linear system (1), with a Luenberger observer (Figure 4),

$$\begin{aligned} \dot{\zeta} &= F\zeta + G_1 y + G_2 u \\ K\hat{x} &= H\zeta + Jy \end{aligned} \tag{14}$$

The observer design equations are known to be[10]

$$\begin{aligned} VA - FV &= G_1 C_2 \\ G_2 &= VB_2 \\ JC_2 + HV &= K \end{aligned} \tag{15}$$

Theorem 3

The closed-loop transfer function matrix can be expressed as

$$T(s) = T_{SF}(s) + \tilde{F}(s)\tilde{M}(s) \tag{16}$$

where

$$T_{SF}(s) = (C_1 - D_{12}K)\,(sI - A + B_2K)^{-1}\,B_1 + D_{11}$$

$$\tilde{F}(s) = (C_1 - D_{12}K)(sI-A+B_2K)^{-1}B_2 + D_{12} \tag{17}$$

and $\tilde{M}(s) = H(sI-F)^{-1}(VB_1-G_1D_{21}) - JD_{21}$

The above theorem is an extension of Theorem 1 ($F(s) = \tilde{F}(s)K$) as it can be seen by letting $V=I$, $J=0$, $H=K$, $G_2=B_2$, and $G_1=L$.

2.4 Exact LTR

An immediate application of Theorem 3 is that exact loop transfer recovery can be achieved when $M(s) = 0$ (Figure 5 and Figure 3). More specifically, via (17), we have

Theorem 4

The necessary and sufficient conditions for exact LTR are those conditions given in (15) and

$$\begin{aligned} VB_1 &= G_1D_{21} \\ JD_{21} &= 0 \end{aligned} \tag{18}$$

The existence of the observer gain matrix G_1 requires the condition $p > q$ be satisfied. The observer design can be accomplished using the following procedures which are devised from a recent result in precise LTR [11]

Theorem 5 ($p > q$, $D_{21} \neq 0$)

(i) perform a QR decomposition of D_{21}, $[Q,R] = qr(D_{21})$

$$Q = [Q_1\ Q_2]\ ,\ R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \tag{19}$$

(ii) Let

$$E = Q^T C_2 = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \tag{20}$$

(iii) Solve the Lyapunov equation

$$V(A-B_1R_1^{-1}E_1) - FV = L_2E_2 \tag{21}$$

where L_2 is any matrix of proper dimension.

(iv) Let

$$G_1 = [VB_1R_1^{-1} \quad L_2]Q^T \tag{22}$$

Given K and V, J and H can be computed, if $r = n + q - p$

$$[J \ H] = [K \ 0] \begin{bmatrix} C_2 & D_{21} \\ V & 0 \end{bmatrix}^{-1} \tag{23}$$

<u>Theorem</u> <u>6</u> ($p > q$, $D_{21} = 0$)

(i) perform a QR decomposition of B_1,

$$[W \ S] = qr(B_1)$$

$$W = [W_1 \quad W_2] \ , \ S = \begin{bmatrix} S_1 \\ 0 \end{bmatrix} \tag{24}$$

(ii) Let

$$\hat{C}_1 = C_2W_1, \ \hat{A}_1 = W_2^TAW_1, \text{ and } \hat{A}_2 = W_2^TAW_2 \tag{25}$$

(iii) perform a QR decomposition of $\hat{C}_1$

$$[\hat{Q}, \ \hat{R}] = qr \ (\hat{C}_1)$$

$$\hat{Q} = [\hat{Q}_1, \ \hat{Q}_2] \ , \ \hat{R} = \begin{bmatrix} \hat{R}_1 \\ 0 \end{bmatrix} \tag{26}$$

and let

$$\hat{E} = \hat{Q}^T \ C_2W_2 \triangleq \begin{bmatrix} \hat{E}_1 \\ \hat{E}_2 \end{bmatrix} \tag{27}$$

(iv) Solve the Lyapunov equation

$$Z\,(\hat{A}_2 - \hat{A}_1\,\hat{R}_1^{-1}\hat{E}_1) - FZ = \hat{L}_2\hat{E}_2 \tag{28}$$

where $\hat{L}_2$ is any matrix of proper dimension.

(v) Set

$$G_1 = [Z\hat{A}_1\hat{R}_1^{-1} \quad \hat{L}_2]\,\hat{Q}^T$$

$$V = ZW_2^T \qquad G_2 = VB_2 \tag{29}$$

$$[J\ H] = K\begin{bmatrix} C_2 \\ V \end{bmatrix}^{-1}$$

Numerical examples to illustrate the above design procedures are presented in Section IV.

2.5 Controller Design via Model Matching

As mentioned previously in this report, our H_∞ design formulation can be viewed as a stable one-sided model matching problem (MMP). Recent research on standard H_∞ control design has resulted in various algorithms for solving model matching problems. It is felt that the method proposed by Hung is more useful to this research.[12]

Since $T_{SF}(s)$ and $\tilde{F}(s)$ are known after the full-state feedback gain K is designed, we can determine $\tilde{M}(s)$ via an appropriate MMP algorithm.

Theorem 7

Let $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\}$ be a minimal realization of $\tilde{M}(s)$, then the observer design equations are

$$
\begin{aligned}
VA-\tilde{A}V &= G_1C_2 \\
JC_2 + \tilde{C}V &= K \\
VB_1 - G_1D_{21} &= \tilde{B} \\
JD_{21} &= -\tilde{D}
\end{aligned}
\tag{30}
$$

From the definition of $\tilde{M}(s)$, the observer parameters F and H are $\tilde{A}$ and $\tilde{C}$, respectively. Other observer parameters J, G_1 and G_2 can be similarly computed as proposed in <u>Theorem 6</u>. The detailed algorithm has yet to be worked out. It should be mentioned here that for the existence of solutions to the design equations (30), the following conditions must be satisfied,

$$
\begin{aligned}
& p > q \\
\text{and} \quad & r(m+p-q) \geq m(n+q-p)
\end{aligned}
\tag{31}
$$

III. Notations and Block Diagrams

3.1 Matrices

			G_1	:	$r \times p$
A	:	$n \times n$	G_2	:	$r \times m$
B_1	:	$n \times q$	H	:	$m \times r$
B_2	:	$n \times m$	J	:	$m \times p$
C_1	:	$\ell \times n$	V	:	$r \times n$
C_2	:	$p \times n$	$T(s)$	:	$\ell \times q$
D_{11}	:	$\ell \times q$	$T_{SF}(s)$	:	$\ell \times q$
D_{12}	:	$\ell \times m$	$F(s)$	:	$\ell \times n$
D_{21}	:	$p \times q$	$\tilde{F}(s)$	:	$\ell \times m$
K	:	$m \times n$	$M(s)$	:	$n \times q$
L	:	$n \times p$	$\tilde{M}(s)$	:	$m \times q$
X_∞	:	$n \times n$	R	:	$p \times q$

Z_∞	:	$n \times n$	R_1	:	$q \times q$
F	:	$r \times r$	Q	:	$p \times p$
Q_1	:	$p \times q$	$\hat{Q}$	:	$p \times p$
Q_2	:	$p \times (p-q)$	$\hat{Q}_1$	:	$p \times q$
E	:	$p \times n$	$\hat{Q}_2$	:	$p \times (p-q)$
E_1	:	$q \times n$	$\hat{R}$	:	$p \times q$
L_2	:	$r \times (p-q)$	$\hat{R}_1$	:	$q \times q$
W	:	$n \times n$	$\tilde{A}$	:	$r \times r$
S	:	$n \times q$	$\tilde{B}$	:	$r \times q$
S_1	:	$q \times q$	$\tilde{C}$	:	$m \times r$
W_1	:	$n \times q$	$\tilde{D}$	:	$m \times q$
W_2	:	$n \times (n-q)$	$\hat{E}$	:	$p \times (n-q)$
$\hat{C}$	:	$p \times q$	$\hat{E}_1$	:	$q \times (n-q)$
$\hat{A}_1$	:	$(n-q) \times q$	$\hat{E}_2$	:	$(p-q) \times (n-q)$
$\hat{A}_2$	:	$(n-q) \times (n-q)$	Z	:	$r \times (n-q)$

3.2 Block Diagrams

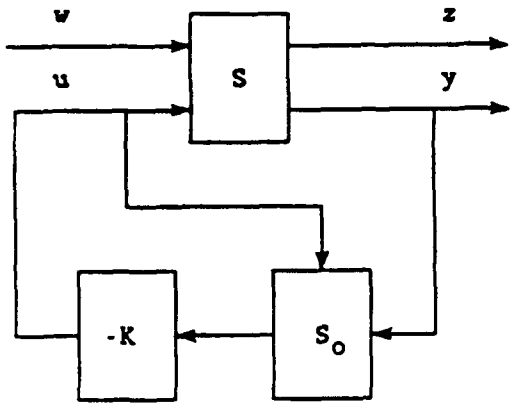


Figure 1. Observer-Based Compensator

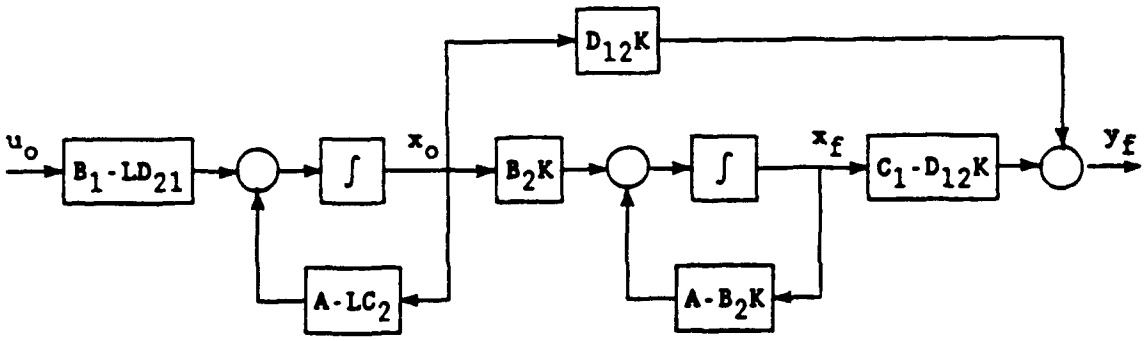


Figure 2. Frequency-Shaped State Estimator

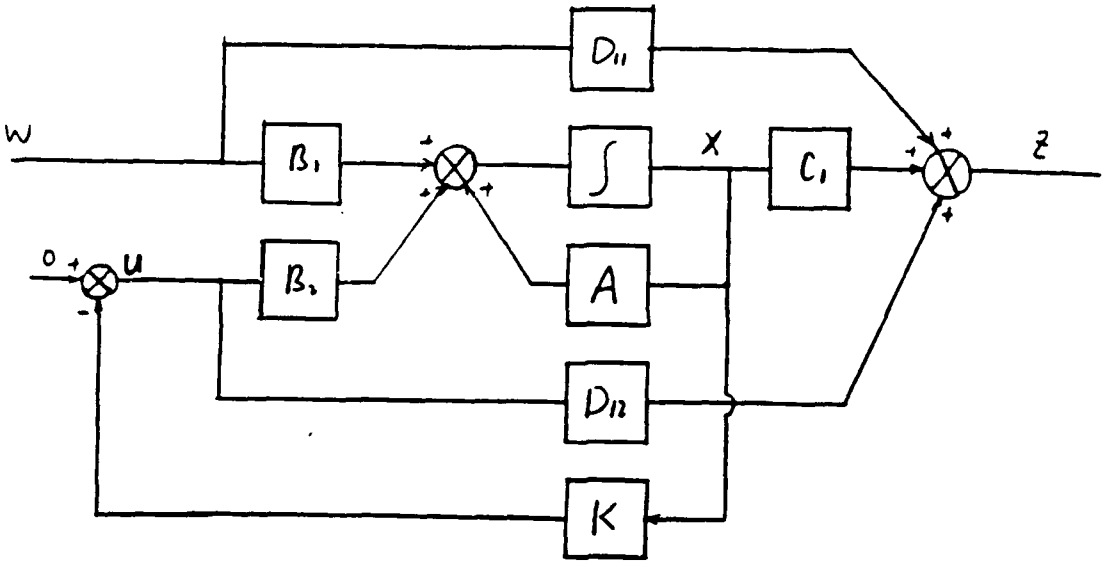


Figure 3. The control system when full states are available

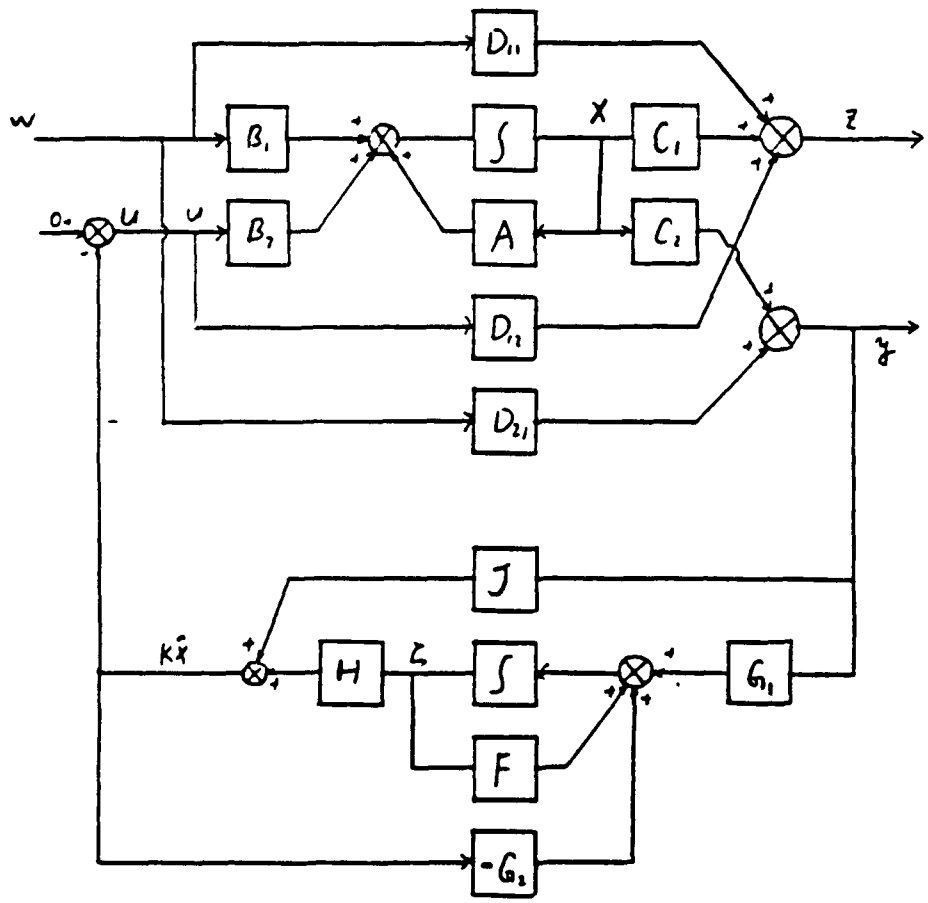


Figure 4. The control system based on observer.

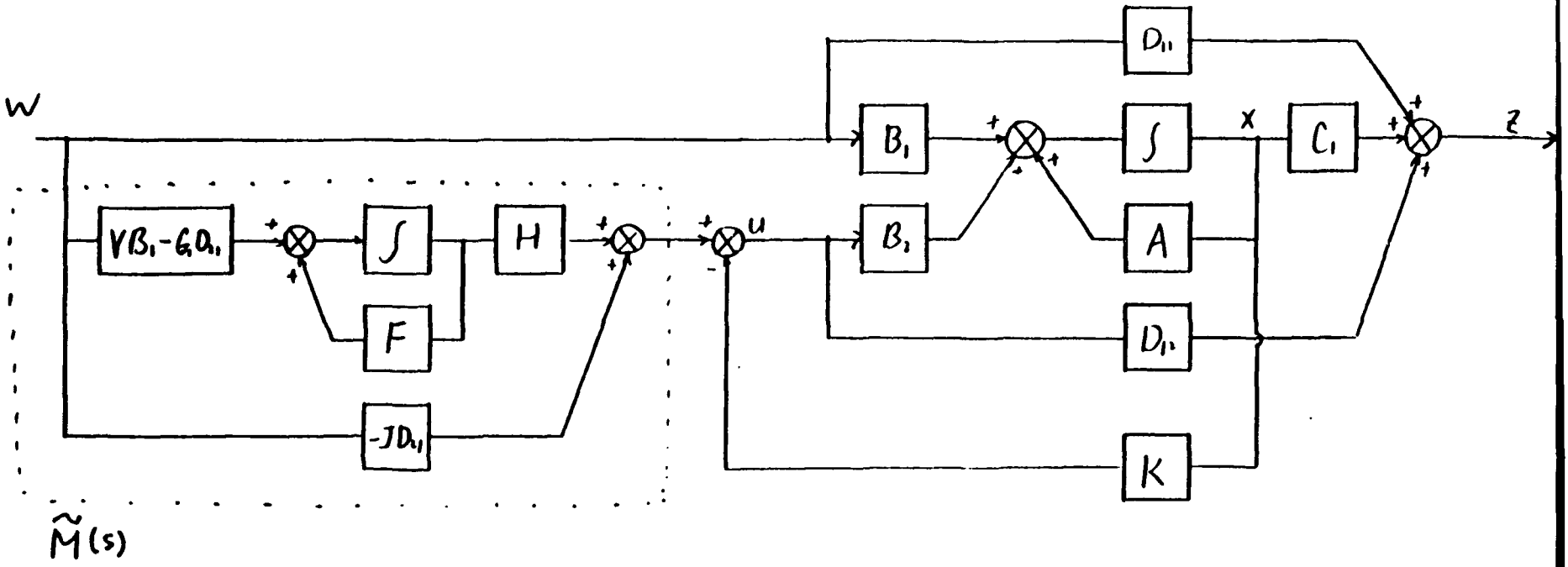


Figure 5. The control system equivalent to that in Figure 4. (For exact LTR, Figure 5 is reduced to Figure 3)

IV. Simulation Results

In this section, we shall present three numerical examples. The first two examples are for exact LTR with $D_{21} \neq 0$ and $D_{21} = 0$. The second example is nominal linearized longitudinal dynamics of the A4D aircraft at flight condition 0.9 Mach and 15,000 ft. altitude.[13] The third example is for approximate LTR covered in subsection 2.1 and 2.2.

4.1 Example 1 (exact LTR, $D_{21} \neq 0$)

<model data>:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 6 & -11 & 6 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$C_1 = [\ 1 \quad 3 \quad 1\ ], \quad C_2 = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$

$$D_{21} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad D_{12} = 2, \quad D_{11} = 1$$

<dimension>:

$n=3,\ q=1,\ m=1,\ \ell=1,\ p=2,\ r=2$

<selection>:

open loop poles : 1, 2, 3
observer poles : -5, -3
free matrix L_2 : $\begin{bmatrix} 10 \\ -10 \end{bmatrix}$, $F = \begin{bmatrix} -5 & 2 \\ 0 & -3 \end{bmatrix}$

<result>:

full state feedback: $K = [9 \quad -4.5 \quad 10.5]$

QR factorization: $Q_1 = \begin{bmatrix} -0.4472 \\ -0.8944 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} -0.8944 \\ 0.4472 \end{bmatrix}$

$R_1 = -2.2361$

$E = Q^T C_2$: $E_1 = [-0.4472 \quad -1.7889 \quad -0.8944]$

$E_2 = [-0.8944 \quad -1.3416 \quad 0.4472]$

Solution to Lyapunov equation:

$$V = \begin{bmatrix} -.2179 & 0.3538 & 0.3278 \\ 4.8291 & 5.3540 & -0.4199 \end{bmatrix}$$

observer parameters:

$$G_1 = \begin{bmatrix} -8.7808 & 4.7991 \\ 11.9677 & 1.5747 \end{bmatrix}$$

$$G_2 = \begin{bmatrix} 0.3278 \\ -0.4199 \end{bmatrix}$$

$$J = [-30.5069 \quad 15.2535]$$

$$H = [-4.2680 \quad 7.9885]$$

<closed loop transfer function>:

with full state feedback, $T(s) = \dfrac{s^3 - 8.5\, s^2 + 336\, s - 193.5}{(s+2)\,(s+1.5)\,(s+1)}$

with reduced-order observer,

$$T(s) = \frac{(s+5)\,(s+3)\,(s^3 - 8.5\, s^2 + 336\, s - 193.5)}{(s+5)\,(s+3)\,(s+2)\,(s+1.5)\,(s+1)}$$

<remark>:

The exact LTR for $D_{21} \neq 0$ can be achieved using a reduced-order observer.

4.2 <u>Example 2</u> (exact LTR, $D_{21} = 0$)

<model data>:

$$A = \begin{bmatrix} -0.0605 & -32.37 & 0 & 32.2 \\ -0.00014 & -1.475 & 1 & 0 \\ -0.0111 & -34.72 & -2.793 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} 0 \\ -0.1064 \\ -33.8 \\ 0 \end{bmatrix}, \quad C_1 = [0 \quad 1 \quad 0 \quad 0], \quad C_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D_{11} = 0, \; D_{12} = 0, \; D_{21} = 0$$

<dimension>:

$n = 4, \quad q=1, \quad m=1, \quad \ell=1, \quad p=2, \quad r=2$

<selection>:

observer poles: -4, -5

free matrix $\hat{L}_2$: $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $F = \begin{bmatrix} -4 & 0 \\ 0 & -5 \end{bmatrix}$

<result>:

full state feedback:

$K = [-1.06312, \; 16.83668, \; -1.38415, \; -32.47965]$

QR factorization of B_1:

$$W_1 = \begin{bmatrix} 0 \\ 0 \\ -1 \\ 0 \end{bmatrix} \qquad W_2 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad S = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\hat{C}_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad \hat{A}_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \hat{A}_2 = \begin{bmatrix} -1.475 & 0.00014 & 0 \\ 32.37 & -0.0605 & -32.2 \\ 0 & 0 & 0 \end{bmatrix}$$

QR factorization of $\hat{C}_1$:

$$\hat{Q}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \hat{Q}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \hat{R}_1 = -1$$

$$\hat{E} = \hat{Q}^T C_2 W_2 : \qquad \hat{E}_1 = [0 \;\; 0 \;\; 0]$$

$$\hat{E}_2 = [0 \;\; -1 \;\; 0]$$

Solution of Lyapunov equation:

$$Z = \begin{bmatrix} 3.25565 & -0.2540 & -2.0443 \\ 1.8596 & -0.2025 & -1.3041 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.2540 & 3.2557 & 0 & -2.0443 \\ 0.2025 & 1.8596 & 0 & -1.3041 \end{bmatrix}$$

Observer parameters:

$$G_1 = \begin{bmatrix} 1 & 1.2113 \\ 1 & 0.5555 \end{bmatrix}$$

$$G_2 = \begin{bmatrix} -0.3464 \\ -0.1979 \end{bmatrix}$$

$$J = [-11.6016 \quad -1.3842]$$

$$H = [-86.5510 \quad 160.5833]$$

<closed loop transfer function>:

with full state feedback:

$$T(s) = \frac{0.8527\ (s-4.9126)\ (s+0.9205)}{(s+30.2167)\ (s+16.3442)\ [(s+1.3802)^2 + 1.2147^2]}$$

with reduced-order observer:

$$T(s) = \frac{0.8527\ (s+4)\ (s+5)\ (s-4.9126)\ (s+0.9205)}{(s+4)\ (s+5)\ (s+30.2167)\ (s+16.3442)\ [(s+1.3802)^2 + 1.2147^2]}$$

<remark>:

The exact LTR for $D_{21}=0$ can be achieved using a reduced-order observer.

4.3 Example 3 (Approximate LTR)

<model data>:[14]

$$A = \begin{bmatrix} -14 & 100 & 0 \\ 1 & 0 & 0 \\ 0 & -10 & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 0 & 0 & 0.5 \\ 0 & 0 & 0 \\ \cdots & \cdots & \cdots \\ 0 & 0 & 1 \\ 1 & 5 & 0 \end{bmatrix} = \begin{bmatrix} C_1 \\ \cdots \\ C_2 \end{bmatrix}$$

$$B = \left[\begin{array}{cccc:c} 0 & 0.1 & 0 & 0 & 100 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{array}\right] \qquad D = \left[\begin{array}{cccc:c} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right] = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$$

<dimension>:

$n=3,\ q=4,\ m=1,\ \ell=2,\ p=2$

<design>:

1. Check 6 initial assumptions.
2. Choose a γ and δ pair.
3. Compute $X_\infty>0 \rightarrow K$.
4. Compute $Z_\infty>0 \rightarrow L$.
5. Check for system stability ($\mathrm{Re}\lambda(A-LC_2)<0$).
6. Compute H_∞-norm, $\|T(s)-T_{sf}(s)\|_\infty$.
7. If any condition fails, repeat step 2- 6.

<system configuration> Figure 1 and Figure 2 of the next page.

<remark>:

For a more complete analysis, (γ=1, δ=1) and (γ=1, δ=0.548) were chosen. Figures 12-15 are the corresponding sigma-plots for $T_f(s)=T(s)-T_{sf}(s)$ and $T_{zw}(s)$. The trade-off between attenuation and bandwidth is apparent. The BW for (γ=1, δ=1) is ≈8 rad/s and for (γ=1, δ=0.548), ≈2000 rad/s. The corresponding step responses are given in Figures 17-20. the step was applied to the y_{1c} command reference input while the other exogeneous inputs were kept at zero. The design intent is for y_{10} to track the command reference input. According to Figures 19-20, the design objective is achieved with $\tau_{63\%}$ =0.996 sec. for (γ=1, δ=1), and $\tau_{63\%}$ =0.711 sec. for (γ=1, δ=0.548). The latter parameter pair provides a performance close to the design found in the paper, $\tau_{63\%}$ =0.6 sec. The output z_1 is the accumulated error e_{ac}. For Figures 17 and 18, e_{ac}≈0.52 and e_{ac}≈0.36, respectively. These plots are in direct casual agreement with the output step responses y_{10} (Figures 19-20). That is, a quicker response will produce less accumulated error (over-damped case). The output z_2 is the control vector u. The negative value indicates the controller is trying to drive the output z_1 to zero; however, this will never happen since there will always be accumulated error.

Eigenvalues:

	$\lambda(T_{zw})$	$\lambda(A)$	$\lambda(A-LC_2)$
γ=1, δ=1→	-19.26	0	-19.21
	-19.21	5.21	-8.02
	-4.42±2.63 j	-19.21	-1.06
	-8.02		
	-1.06		

	$\lambda(T_{zw})$	$\lambda(A)$	$\lambda(A-LC_2)$
γ=1, δ=0.548→	-2407.7	0	-2407.7
	-19.21	5.21	-19.21
	-19.26	-19.21	-1.63
	-4.42±2.63 j		
	-1.63		

Notice that one of the poles is weakly observable, $\lambda_3(A)$=-19.21. The pole is

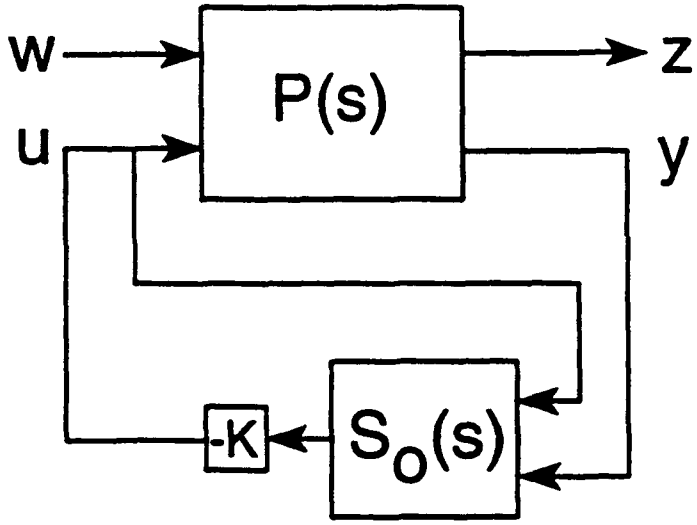


Figure 1. General block diagram: Plant P(s) augmented with observer controller $S_O(s)$ and observer state gain K.

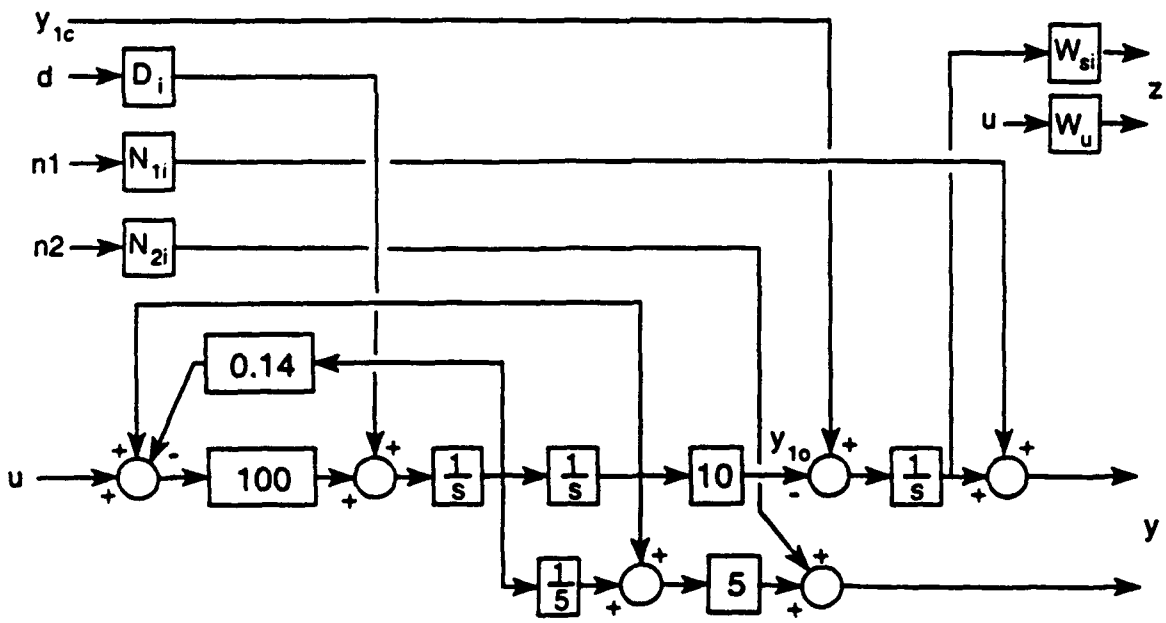


Figure 2: System1 block diagram: Original plant augmented with weights D_i, N_{1i}, N_{2i}, W_{si}, W_u, and an integrator placed at output y_1.

already quite stable and therefore, is not moved by the controller.

<results>

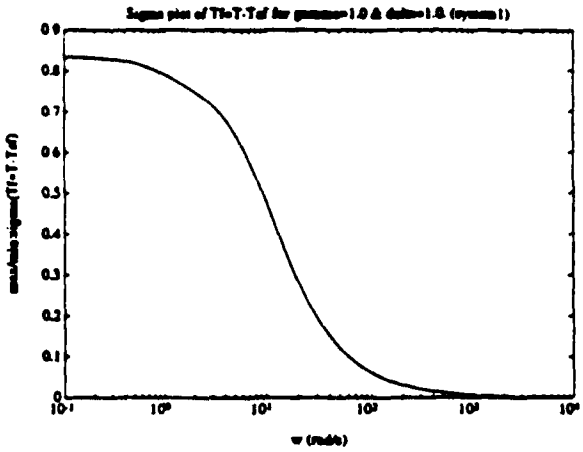


Figure 12

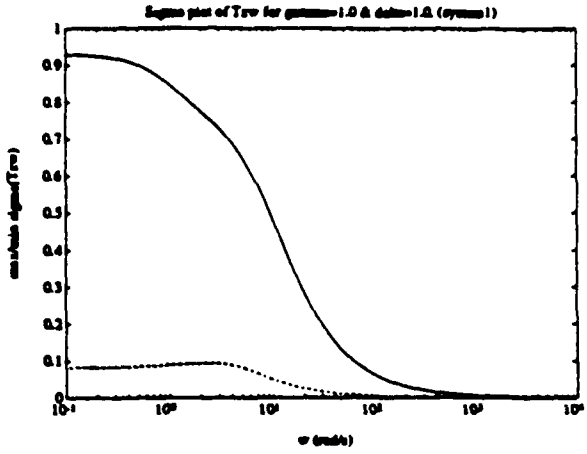


Figure 13

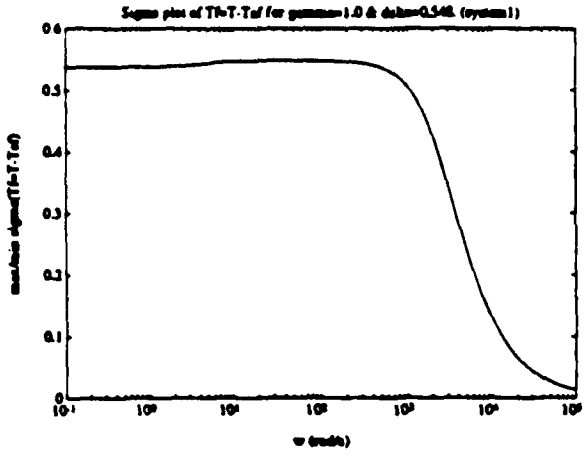


Figure 14

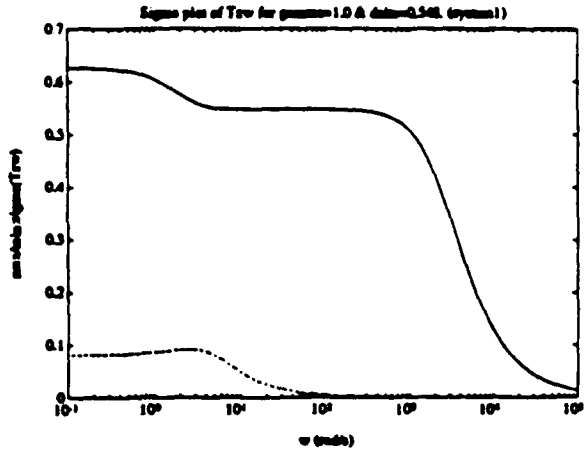


Figure 15

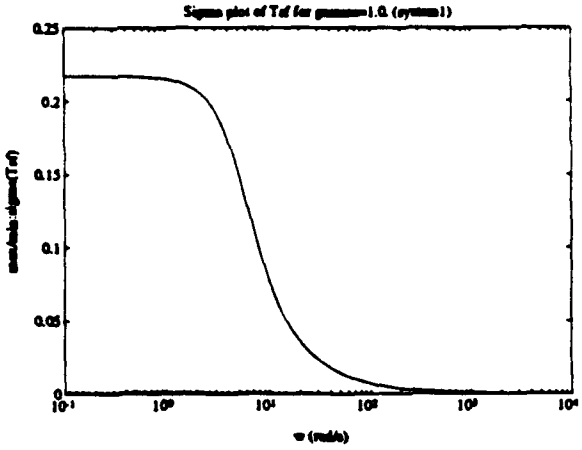


Figure 16

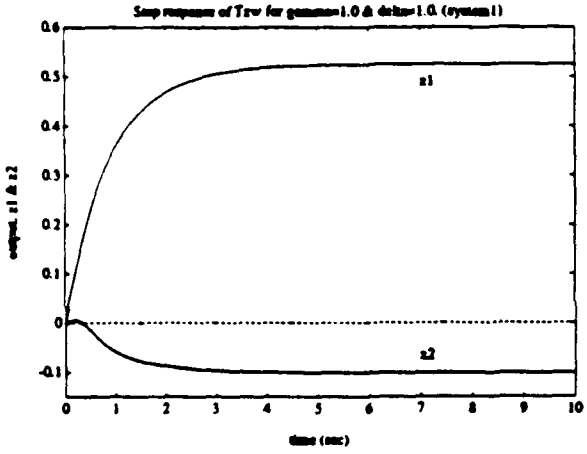


Figure 17

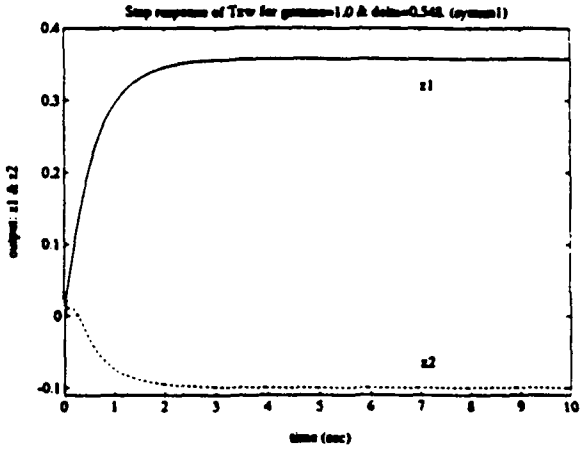


Figure 18

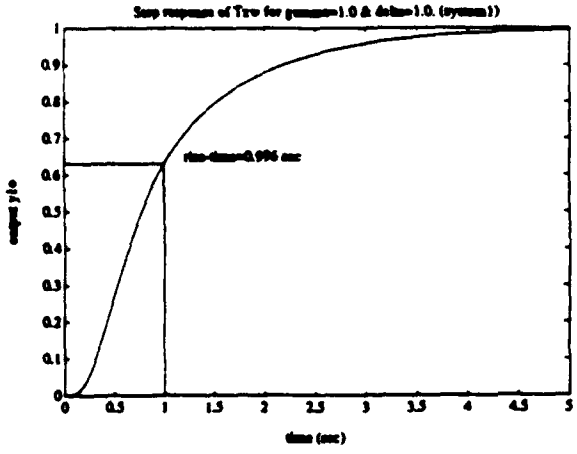


Figure 19

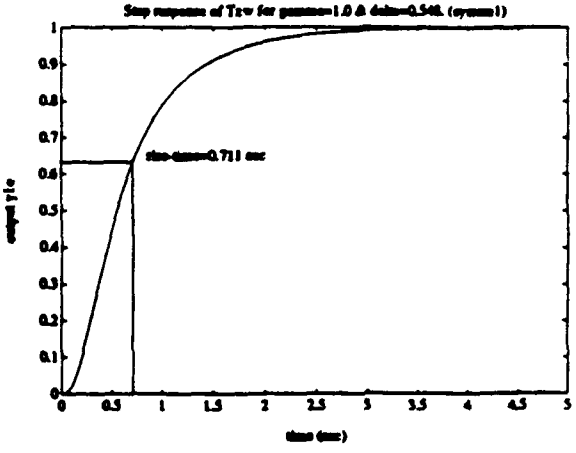


Figure 20

```
gamma= 1.0000
delta= 1.0000

X=
1.411e-03        2.969e-02        -5.097e-03
2.969e-02        6.416e-01        -1.428e-01
-5.097e-03       -1.428e-01       9.893e-02

lambda(X)=
3.281e-06 +0.000e+00j
6.782e-01 +0.000e+00j
6.374e-02 +0.000e+00j

Z=
3.288e+00        6.314e-01        -1.269e+00
6.314e-01        1.213e-01        -2.437e-01
-1.269e+00       -2.437e-01       1.652e+00

lambda(Z)=
4.093e+00 +0.000e+00j
9.681e-01 +0.000e+00j
9.260e-06 +0.000e+00j

K=
1.411e-01        2.969e+00        -5.097e-01

L=
-1.269e+00       6.445e+00
-2.437e-01       1.238e+00
1.652e+00        -2.488e+00

lamba(A-LC2)=
-1.921e+01 +0.000e+00i
-8.016e+00 +0.000e+00i
-1.063e+00 +0.000e+00i

||T-Tsf||= 8.344335e-01  @ w= 0.000

Design Status:
X positive definite:              passed.
Z positive definite:              passed.
Feedback system stable:           passed.
Feedback system norm < delta:     passed.
```

```
gamma= 1.0000
delta= 0.5480

X=
1.411e-03        2.969e-02        -5.097e-03
2.969e-02        6.416e-01        -1.428e-01
-5.097e-03       -1.428e-01       9.893e-02

lambda(X)=
3.281e-06 +0.000e+00j
6.782e-01 +0.000e+00j
6.374e-02 +0.000e+00j

Z=
5.114e+02        9.822e+01        -4.778e+02
9.822e+01        1.887e+01        -9.178e+01
-4.778e+02       -9.178e+01       4.492e+02

lambda(Z)=
9.780e+02 +0.000e+00j
1.477e+00 +0.000e+00j
9.261e-06 +0.000e+00j

K=
1.411e-01        2.969e+00        -5.097e-01

L=
-4.778e+02       1.003e+03
-9.178e+01       1.926e+02
4.492e+02        -9.367e+02

lamba(A-LC2)=
-2.408e+03 +0.000e+00i
-1.921e+01 +0.000e+00i
-1.633e+00 +0.000e+00i

||T-Tsf||= 5.474570e-01  @ w= 106.525

Design Status:
X positive definite:              passed.
Z positive definite:              passed.
Feedback system stable:           passed.
Feedback system norm < delta:     passed.
```

V. Conclusions and Recommendations for Future Research

The main task of the described research under the present contract has been to develop a new approach for H_∞ control design. The ultimate objective is to devise H_∞ theory-oriented numerical algorithms which would offer reduced-order H_∞ compensators.

In this report, we have presented two H_∞ design results. The first result is for approximate H_∞ loop transfer recovery based on full-order observers, while the second result is for exact (precise) LTR based on reduced-order observers. Three numerical examples are also included to demonstrate the use of the presented design procedures. It is mentioned here that the two results are not meant to be used for the same control design problem, since the technical assumptions leading to the two results are different.

Several research topics are recommended for future research:

1. Use the notion of 'robust observer design' to extend the presented results for control systems with parametric uncertainties.

2. Revise the existing results in H_∞ state feedback design so that the obtained gain matrix K can be used directly for exact LTR via reduced-order observers.

3. Perform more computer simulations using practical control examples to assess the numerical aspects of the proposed design algorithms.

4. Compare the proposed exact H_∞/LTR method with Niemann et. al. who formulated the H_∞/LTR as a singular H_∞ control problem.

References

1. Doyle, J.C., K. Glover, P.P. Khargonekar and B.A. Francis, "State-Space Solutions to Standard H_2 and H_∞ Control Problems," IEEE Trans. Automat. Control, Vol. AC-34, 1989, pp. 831-847.

2. H. Kimura, Y. Lu and R. Kawatani, "On the Structure of H_∞ Control Systems and Related Extensions," IEEE Trans. Auto. Contr., Vol. AC-36, No. 6, 1991, pp. 653-667.

3. I. Yaesh and U. Shaked, "A Transfer Function Approach to the Problems of Discrete-Time Systems: H_∞-Optimal Linear Control and Filtering, "IEEE Trans. Auto. Contr., Vol, 36, No. 11, 1991, pp. 1264-1271.

4. M. Fujita, K. Uchida and F. Matsumura, "Asymptotic H_∞ Disturbance Attenuation Based on Perfect Observation," IEEE Trans. AC, Vol 36, No. 7, 1991, pp 875-880.

5. H. H. Yeh, S.S. Banda, A.G. Sparks and D.B. Ridgely, "Loop Shaping in Mixed H_2 and H_∞ Optimal Control," Proceedings 1991 American Control Conference, Boston, Mass., pp. 1165-1170.

6. J. Rawson, C.S. Hsu, H.H. Yeh and S.S. Banda "H_∞/LTR: A Loop Shaping Method for H_∞ Output Feedback Compensator Design," Proceedings 1991, American Control Conference, Boston, Mass., pp. 2196-2201.

7. H.H. Niemann, Per Sogaard-Andersen and J. Stoustrup, "Loop Transfer Recovery for General Observer Architectures," International Journal of Control, Vol., 53, No. 5, 1991, pp. 1177-1203.

8. Khargonekar, P.P. I.R. Petersen and M.A. Rotea, "H_∞ Optimal Control with State Feedback, "IEEE Trans. Auto. Contr., Vol. AC 33, 1988, pp. 786-788.

9. C.S Hsu and J.L. Rawson, "H_∞ Design Based on Loop Transfer Recovery and Loop Shaping," Final Report, 1990, USAF-UES SFRP/GSRP. Contract No: F49620-88-C-0053.

10. J. O'Reilly, Observers for Linear Systems, London, Academic Press, 1983.

11. M.M. Monahemi, J.B. Barlow, and D.P. O'Leary, "The Design of Reduced-Order Luenberger Observers with Precise LTR," Proceeding Guidance, Navigation and Control Conference, 1991.

12. Y.S. Hung, "H_∞ Optimal Control, Part 1. Model Matching," International Journal of Control, Vol. 49, No. 4, 1989, pp. 1291-1330.

13. I.R. Petersen and C.V. Hollot, "High Gain Observers Applied to Problems in the Stabilization of Uncertain Linear Systems, Disturbance Attenuation and H_∞ Optimization," Int'l J. Adaptive Control and Signal Processing, Vol 2, 1989. pp. 347-369.

14. W.S. Levine and R.T. Reichert, "An Introduction to H_∞ Control System Design," Proceedings 20th Conference on Decision and Control, Honolulu, Hawaii, 1990, pp. 2966-2974.

Appendix: Derivation of Main Results

We shall present the derivation of Theorem 2, Theorem 3 and Theorem 5 in the sequel. Theorem 1 is a special case of Theorem 2, and hence omitted. Theorem 4 follows Theorem 3, while Theorem 6 is very similar to Theorem 5.

Appendix A (Proof of Theorem 2)

<u>Proof of Theorem 2</u>: Figure 2 gives a realization of the transfer function matrix $T(s) - T_{sf}(s) = F(s)\Phi_0(s)(B_1 - LD_{21})$. A state-space description is:

$$\begin{bmatrix}\dot{x}_o(t)\\ \dot{x}_f(t)\end{bmatrix} = F\begin{bmatrix}x_o(t)\\ x_f(t)\end{bmatrix} + G\,u_o(t)$$

$$y_f(t) = H\begin{bmatrix}x_o(t)\\ x_f(t)\end{bmatrix}$$

where, $F = \begin{bmatrix}A - LC_2 & 0\\ B_2K & A - B_2K\end{bmatrix}$, $G = \begin{bmatrix}B_1 - LD_{21}\\ 0\end{bmatrix}$ and $H = [D_{12}K \quad C_1 - D_{12}K]$.

The problem is to find L such that H_∞ norm of the transfer function between u_o and y_f is less than δ. This norm bound will be satisfied if F is stable and there exists an $X \geq 0$ such that :

$$U = F^TX + XF + \delta^{-2}XGG^TX + H^TH \leq 0.$$

Let X = block diag$\{X_1, X_\infty\}$ and partition U into 4 n×n blocks:

$$U = \begin{bmatrix}U_{11} & U_{12}\\ U_{21} & U_{22}\end{bmatrix}.$$

where,

$$\begin{aligned}U_{22} &= X_\infty(A - B_2K) + (A - B_2K)^TX_\infty + C_1^TC_1 + X_\infty B_2(D_{12}^TD_{12})^{-1}B_2^TX_\infty\\ &= X_\infty A - X_\infty B_2(D_{12}^TD_{12})^{-1}B_2^TX_\infty + A^TX_\infty - X_\infty B_2(D_{12}^TD_{12})^{-1}B_2^TX_\infty\\ &\quad + C_1^TC_1 + X_\infty B_2(D_{12}^TD_{12})^{-1}B_2^TX_\infty\\ &= -\gamma^{-2}X_\infty B_1B_1^TX_\infty\end{aligned}$$

by using the state feedback Riccati equation (3). The off-diagonal

blocks vanish:

$$U_{12} = U_{21} = X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty - X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty = 0$$

Finally, the upper left-hand block is also made to vanish.

$$\begin{aligned} U_{11} &= X_1(A - LC_2) + (A - LC_2)^T X_1 + \delta^{-2} X_1(B_1 B_1^T + LD_{21}D_{21}^T L^T)X_1 \\ &\quad + X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty \\ &= X_1 A + A^T X_1 + \delta^{-2} X_1 B_1 B_1^T X_1 + X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty \\ &\quad + (\delta^{-1} X_1 L D_{21} D_{21}^T - \delta C_2^T)(D_{21}D_{21}^T)^{-1}(\delta^{-1} X_1 L D_{21} D_{21}^T - \delta C_2^T)^T \\ &\quad - \delta^2 C_2^T (D_{21}D_{21}^T)^{-1} C_2 \end{aligned}$$

If

$$\delta^{-1} X_1 L D_{21} D_{21}^T = \delta C_2^T$$

and

$$\begin{aligned} 0 &= X_1 A + A^T X_1 + \delta^{-2} X_1 B_1 B_1^T X_1 + X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty \\ &\quad - \delta^2 C_2^T (D_{21}D_{21}^T)^{-1} C_2 \end{aligned} \qquad (*)$$

then $U_{11} = 0$. If the solution to $(*)$ is $X_1 > 0$, then $X \geq 0$ and we can solve for L as

$$L = \delta^2 X_1^{-1} C_2^T (D_{21}D_{21}^T)^{-1}.$$

Equation $(*)$ can then be rewritten as

$$\begin{aligned} 0 &= AZ_\infty + Z_\infty A^T + Z_\infty(\delta^{-2} X_\infty B_2(D_{12}^T D_{12})^{-1} B_2^T X_\infty - C_2^T (D_{21}D_{21}^T)^{-1} C_2) Z_\infty \\ &\quad + B_1 B_1^T \end{aligned}$$

where,

$$Z_\infty = \delta^2 X_1^{-1}.$$

In this case, L is defined as in equation (13).

Since $U_{22} \leq 0$ and the remaining blocks of U are identically 0, $U \leq 0$. All that remains to show is that F is stable. This follows easily from the fact that $A - B_2K$ and $A - LC_2$ were required to be stable at the state and observer stages of the design procedure respectively.

■

Appendix B (Proof of Theorem 3)

From equations (1), (14) and (15), we have

$$\begin{bmatrix} \dot{x} \\ \dot{\zeta} \end{bmatrix} = \begin{bmatrix} A - B_2JC_2 & -B_2H \\ G_1C_2 - G_2JC_2 & F - G_2H \end{bmatrix} \begin{bmatrix} x \\ \zeta \end{bmatrix} + \begin{bmatrix} B_1 - B_2JD_{21} \\ (G_1 - G_2J)D_{21} \end{bmatrix} w \qquad (B.1)$$

$$z = [C_1 - D_{12}JC_2 \quad -D_{12}H] \begin{bmatrix} x \\ \zeta \end{bmatrix} + (D_{11} - D_{12}JD_{21})\, w \qquad (B.2)$$

Applying a coordinate transformation to (B.1) gives

$$\dot{\eta} = \begin{bmatrix} A - B_2K & -B_2H \\ 0 & F \end{bmatrix} \eta + \begin{bmatrix} B_1 - B_2JD_{21} \\ -VB_1 + G_1D_{21} \end{bmatrix} w \qquad (B.3)$$

$$z = [C_1 - D_{12}K \quad -D_{12}H]\ \eta + D_{11} \quad -D_{12}JD_{21} \tag{B.4}$$

where $\eta = N \begin{bmatrix} x \\ \zeta \end{bmatrix}$,

$$N \triangleq \begin{bmatrix} I & 0 \\ -V & I \end{bmatrix}$$

The closed-loop transfer function matrix T(s) can be obtained from (B.3) and (B.4),

$$T(s) = [C_1 - D_{12}K \quad -D_{12}H] \begin{bmatrix} sI - A + B_2 K & B_2 H \\ 0 & sI - F \end{bmatrix}^{-1} \begin{bmatrix} B_1 & -B_2 JD_{21} \\ -VB_1 & + G_1 D_{21} \end{bmatrix}$$

$$+ D_{11} \quad -D_{12}JD_{21} \tag{B.5}$$

Expanding the RHS of (B.5) and after some algebraic manipulation yields the expected result

$$T(s) = T_{SF}(s) + \tilde{F}(s)\ \tilde{M}(s) \tag{B.6}$$

where $\tilde{F}(s)$ and $\tilde{M}(s)$ are defined as in (17).

<u>Appendix C</u>: (Proof of Theorem 5)

The proof of this theorem can be established by verifying $G_1 D_{21} = VB_1$ and $VA - FV = G_1 C_2$. From (22),

$$G_1 = [VB_1 R_1^{-1} \ L_2]\ Q^T \quad \text{, where Q is an orthogonal matrix,}$$

which implies

$$G_1 D_{21} = [VB_1 R_1^{-1} \ L_2]\ Q^T Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

$$= [VB_1 R_1^{-1} \ L_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

$$= VB_1 \qquad \text{, since } Q^T Q = I$$

Next, from (20) and (21) we have

$$VA - FV$$

$$= VB_1R_1^{-1} E_1 + L_2E_2$$

$$= [VB_1R_1^{-1} \quad L_2] \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}$$

$$= G_1Q \cdot Q^TC_2$$

$$= G_1C_2 \quad , \text{ since } QQ^T = I$$

STRESS WAVE PROPAGATION THROUGH THE THICKNESS OF GRAPHITE/EPOXY LAMINATED PLATES USING PVDF SENSORS

Final Report: December 1991

Principal Investigator: Dr. David Hui
University of New Orleans
Dept of Mechanical Engineering
New Orleans, LA 70148

Tel: (504)-286-6192
FAX: -7413

co-investigators: Dr. Piyush Dutta,
US Army Cold Regions Research
and Engineering Laboratory, NH 03755

Arnold Mayer and Greg Czarnecki
Wright Patterson Air Force Base, OH 45433

funding period: January 1991- December 1991

funding agency: Universal Energy System, UES Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432-1894

attention: Susan Espy,
Assistant Program Manager
Tel: (513)-426-6900
FAX: (513)-429-5413

for Air Force Office of Scientific Research
Research Initiation Program

this is an invited paper presented at the
22nd Midwestern Mechanics Conference, October 6-9, 1991
held at Univ. of Missouri-Rolla

Detailed final report will be published as US Army CRREL report, 1992.

ABSTRACT

The objective of this study is to determine the stress wave (or pulse) propagation through the thickness of a graphite-epoxy laminated plate at room temperature and at low temperature. This is part of an overall study to understand the damage of these plates under normal projectile impact. Upon a sharp impact, the stress wave propagates from the impact point into the rest of the material. It is found that the imbedded polyvinylidene fluoride sensors are able to predict the wave velocities and wave attenuation.

1. INTRODUCTION

The impact behavior of laminated plates is an important topic since composite plates are known to respond to impact loading and energy dissipation in a very different way than metallic plates. In fact, impact resistance is one of the most serious weakness of composite material plates. Excellent recent survey articles on this topic was reported by Abrate (1991) and Cantwell and Morton (1991). The impact of metallic plates by spherical balls were reported by Goldsmith (1960), Goldsmith and Lyman (1960) and more recently by Sondergaard et al (1990).

The concepts of imbedded strain gages were employed by Daniel and Wooh (1985, 1990) to study the deformation and damage of composite laminates under impact loading. The characteristic features of the strain records are associated with specific failure modes of the laminates. The load history, imparted energy and transient strains at various locations through the thickness were obtained. The wave propagation behavior of composites using imbedded gages were not discussed. Wave propagation in transversely impacted composite laminates were obtained by Kim and Moon (1979) and Daniel et al. (1979), but no imbedded sensors were used for the wave propagation studies.

The objective of this work is to examine the wave velocities and wave attenuation in the thickness direction using the PVDF sensors which are imbedded in the interior of the laminate. Upon a sharp impact, the stress wave propagates from the imapct point into the rest of the material. Immediately below the impact point over a small area, the stress wave can be assumed to propagate with a plane front in the thickness direction. A plane pressure sensor of relatively small dimension will respond to this propagating stress wave front. A series of such sensors embedded in various depths in the interfaces of the lamina will respond to this incoming stress pulse in sequence at which the sensor meet the pulse. A measurement of time difference between the start of successive pulses, divided by the distance between the sensors, would give the velocity of the stress wave between sensors.

The present impact problem is concerned with extremely short durati pulse so that the wavelength is short relative to the thickness of the individual lamina. Such short-wavelength pulse is especially needed in laminates which contain relatively few layers since one needs to examine the reflection, transmission and superposition of wave through a laminated plate.

There are three phases involving the impact of laminated plates due to these tiny spherical projectiles. They are (i) drop test of these projectiles which does not cause damage (ii) using a gas gun to shoot these projectiles which will cause partial damage but still no perforation (iii) with perforation. The present work is concerned with the first phase which is necessary to demonstrate the validity of the techniques. Further, the velocity of the stress wave will enable one to measure the Young's modulus in the thickness direction, which is known to be hard to measure. The drop test enables one to determine the compressive stresses of a particular lamina as the wave propagates towards the free back surface and the superposition of the tensile stresses as the wavea reflected from free back surface.

The use of piezoelectric polymer as a material which transforms an electric field to a small mechanical mechanical deformation directly through a readjustment of internal polarization is well known (Cundari and Abedian 1991). The polyvinylidene fluoride (PVDF) piezoelectric sensors are imbedded in the interior of the laminated plate.

The purpose of these sensors is to enable one to "look" inside the composite specimens and to determine the sequence and propagation of the stress waves. Of particular concern is the duration of stress wave as it crosses a lamina due to impact and the reflection of waves from the back free surface. It appears that the data collection Nicolet instruments was sensitive enough to measure the stress pulse through the charge in the PVDF and hence the force applied to the sensor within the laminate. The ultimate goal is to understand the "process" , rather than the final stage, of damage due to the reflection and refraction of these waves causing delamination, matrix cracking and fiber breakage.

The wavelength of the pulse is assumed to be short relative to the lamina thickness but long relative to the diameter of the individual fiber. Thus, the material is governed by the effective properties of the equivalent homogeneous material. The "interface" effects are neglected and the analysis is thus, identical to that of the homogeneous material.

The laminated plate consists of four sets of layers (each set consists of seven layers). The plate dimension is 4" by 4". The plate is clamped in a fixture so as to becomes a circular plate with a diameter of three inches. The fixture was used in previous experiments involving Hopkinson bar tests (see Altamirano 1991).

2. WAVE SPEED AND ATTENUATION

As a first approximation, the "interface" effects are neglected and the composite plates can be modelled as homogeneous isotropic material. The longitudinal wave speed is (Kolsky 1963)

$$v_L = \left((gE/\rho)(1-\nu)/\left[(1+\nu)(1-2\nu)\right]\right)^{1/2}$$

and the two transverse wave speeds are identical and they are

$$v_T = \left((gE/\rho)(1/2)/(1+\nu)\right)^{1/2}$$

In the above, E is Young's modulus, is Poisson's ratio and is the density of the material. Note that the above velocities are independent of the frequency so that for plane waves, a pulse shape composed of a spectrum of frequencies, can propagate without distortion of its shape. The laminae are manufactured by Fiberite (see Table 1).

Using the following material properties,

E_t = transverse compressive modulus = 1.2 x10**6 psi
ν = Poisson's ratio = 0.25
ρ = density = 0.055 lb-mass/in**3
g = gravity = 386 in/sec**2

one obtains the longitudinal wave speed c_L = 8382 ft/sec.

The experimental wave speed was measured by any two of the three imbedded sensors. The wave speed, measured by two sensors which are seven plies apart and the duration of 0.5 microsecond (that is, 500 nanoseconds) found from

c_L = lamina thickness/propagation time = (0.041/12)/(500x10**-9)
= 6833 ft/sec

It appears that the presence of the additional epoxy in the PVDF sensors causes the wave to travel at a slower speed. The measured velocity corresponds to the transverse modulus of 0.8x10**6 psi.

The wave propagation is accompanied by attenuation of its amplitude for three major effects (i) geometric attenuation, (ii) interfacial friction between fibers and matrix and (iii) the interlaminar friction between adjacent plies. Geometric attenuation is due to the spreading out of the wave in a spherical direction starting from the point of impact. Normally, the geometric attenuation is predominant over the remaining effects. However, since the distance of wave propagation in the thickness direction is so small compared to the planar direction, it is not necessarily the dominant factor.

The dispersion effects are expected to be small since all layers (with different stacking orientation) are identicla in the thickness direction. Careful calibration of the PVDF sensor is needed. Since the force on the PVDF sensors are proportional to the charge, the proportional constants are determined which are found to be a function of the applied force.

The analysis is confined to the transient stage since one is interested to determine the process of damage in the first crucual nano-seconds. The subsequent vibration problem in the transition from transient to steady state stage is an important problem. Since one is interested in the initial process of damage, particular emphpasis is placed on the wave velocities and attenuation in the transient stage. The subsequent steady state vibration problem is also of interest as it gives the complete damping process involving wave attenuation.

3. EXPERIMENTAL WORK

The experimental setup for the drop test is shown in Figure 1. The sample square plate is clamped in a fixture to produce a circular plate with diameter being three inches. The plies and orientations of the laminate are

$$(0_7, 45_7, 90_7, 0_7)$$

A schematic diagram for the impactor and the clamped circular plate typical stress versus time curves are shown in Figure 2.

The PVDF sensors of 28 micron (1 micron= 10**-6 meter) thickness are imbedded at every seven plies interval where adjacent plies changes the orientation. Two thin lead wires of diameter 41 standard gauge are soldered onto each PVDF sensor and extended out to the edge of the plate and are connected to the digital oscilloscope. The velocity measurements are made by monitoring the pressure sensed by two consecutive sensors by a high resolution digital oscilloscope capable of sampling at 50 nano-seconds interval. Attenuation and energy of the stress pulses were monitored by a four channel digital oscilloscope having a sampling rate of 500 nano-seconds.

The stress waves are generated by dropping spherical steel balls of six different sizes (0.039", 0.219", 0.250", 0.344", 0.500" and 0.563" diameters). The resulting stress versus time curves for each of these balls are presented in Figure 3. The wave forms are recorded on the high speed digital waveform oscilloscope as shown in Figure 4. Typical waveform is shown in Figure 5a and the repeatability of the waveforms from the PVDF sensors from eight consecutive drop ball impacts are demonstrated in Figure 5b.

4. RESULTS

Figure 6 shows the front part of the plot of two waves generated by the drop of a one millimeter diameter steel ball. The two complete wave forms are also shown in this figure. The time difference between the start of the stress waves shown in Figure 6 is the time that the wave has taken to propagate from the first sensor to the next sensor through a thickness of seven plies (0.041"). The velocity is computed by dividing this distance by the time interval.

Figure 7 shows the amplitude decay of the stress wave as it propagates through the three consecutive snsors located at seven ply intervals. It can be shown that as the wave propagates, the peak amplitude decay. These tests are repeated with six different sizes of steel balls. For each ball, five measurements of the waveform decay are made. This figure depicts the wave form recorded from the drop of a single size ball (1 mm) to demonstrate the repeatability of the tests.

The energy absorption by the plate due to impact can be computed from the difference between the initial kinetic energy and the rebound kinetic energy. The energy absorptino E can be compared with the energy of the stress wave E recorded by the first PVDF sensor using

$$E = (AC/E) \int \sigma^2 \, dt$$

where A is the equivalent affecte area, C is the wave speed, o is the stress, t is time and E is Young's modulus. The discrepancy between E and E is due to the attenuation of the wave which was due to the three effects mentioned above. Such discrepancy for different sizes of the balls will enable one to estimate the equivalent affected area.

An interesting results are obtained by plotting [illegible] duration of the pulse versus the size of the spherical indentor. It can be seen that the experimental data fits the following equation,which can be compared with the equation by Bousar.

Finally the waveforms plot of stress versus time for the three imbedded sensors are shown in Figure 8. One vertical division represents 2.015 Pascal and a horizontal division represents 1.953 micro-sec. The first curve (from the first sensor) is replotted in Figure 9. Note that the curve drops after the first peak which is an indication that the reflected wave from the free back surface has arrive as tensile wave. A re-construction of the superposition of tensile and compressive wave as a function of time can be seen in Figure 9. Finally, the same superposition of compressive and tensile waves as sensed by the second and third sensors are snwon in Figure 10.

5. CONCLUSIONS

The embedded PVDF sensors are fond to be effective to study the wave propagation in the thickness direction of the lamianted plates. The interference from reflected wave can be minimized or even eliminated by reducing the size of the impactor balls. The method provides a way to determine the Young's modulus in the thickness direction from the measured wave speed. This method would also allow an estimation of the wave attenuation in the thickness direction.

6. REFERENCES

Abrate, S. (1991), "Impact on Laminated Composite Materials", ASME Applied Mechanics Review, Volume 44, No. 4, April, pp. 155-190.

Altamirano, M. (1991), "Experimental Investigation of High and Low Impact Energy Absorption of AS4/3502 Graphite/Epoxy Panels" MS thesis, University of New Orleans, May.

Cantwell, W.J. and Morton, J. (1991), "The Impact Resistance of Composite Materials - a review", Composites, Vol. 22, No. 5, September, pp. 347-362.

Cundari, M. and Abedian, B. (1991), "The Dynamic Behjavior of a Polyvinylidene Fluoride Piezoelectric Motional Device" Smart Structures and Materials, edited by G. K. Haritos and A. V. Srinivasan, ASME AD-volume 24, AMD-Volume 123, Winter Annual Meeting, Atlanta, Dec. 1-6, 1991, pp. 25-31.

Daniel, I.M., Liber T. and LaBedz, R.H. (1979), "Wave Propagation in Transversely Impacted Composite Laminates", Experimental Mechanics, Vol. 19, No. 1, pp. 9-16.

Daniel, I.M. and Woon, S.C. (1985), "Embedded Gages for Study of Tansient Deformation and Dynamic Fracture in Composites", Proc. of Fall Meeting of SEM, Grenelefe, Fl, November, pp. 62-68.

Daniel, I.M. and Wooh, S.C. (1990), "Deformation and Damage of Composite Laminates under Impact Loading", Impact Response and Elastodynamics of Composites, edited by A.K. Mal and Y.D.S. Rajapakse, AMD-Volume 116, Winter Annual Meeting, Nov. 25-30, pp. 11-26.

Goldsmith, W. (1960), "Impact: The Theory and Physical Behavior of Colliding Solids", Edward Renold, London.

Goldsmith, W. and Lyman, P.T. (1960), "The Penetration of Hard-Steel Spheres into Plane Metal Surfaces", ASME J. of Applied Mechanics, December, pp. 717-725.

Kim, B.S. and Moon, F.C. (1979), "Impact Induced Stress Waves in an Anisotropic Plate", AIAA Journal, Vol. 17, pp. 1126-1133.

Kolsky, H. (1963), "Stress Waves in Solids" Dover Publications, New York.

Sondergaard, R., Chaney, K. and Brennen, C.E. (1990), "Measurements of Solid Spheres Bouncing Off Flat Plates", ASME J. of Applied Mechanics, Vol. 57, September, pp. 694-699.

Appendix A1 The PVDF Sensors

The PVDF sensor being used has the following properties:

Thickness = 28x10**-6 m = 1.102 x 10**-3 inch)
Lead wires = 41 swg
readout: digital oscilloscope
speed = 50x10**-9 sec/div

Number of channels = 2 (for velocity study)
= 4 (for attenuation study)

Lamina Thickness (seven plies) = 1.04x10**-3 m = 0.041 inch
Sensor to lamina thickness = 1/ 37.2

A schematic diagram of the PVDF film is shown in Figure A1. The PVDF sensor measurement principle is shown in Figure A1. The applied stress σ can be found from

$$\sigma = Q/(A.\ d)$$

where Q is the charge, A is the sensor area and d is the charge sensitivity. Further,

$$d\sigma/dt = (1/(Ad))\ dQ/dt$$

and the current I can be found from

$$dQ/dt = I = V/R$$

where V is voltage and R is resistance. Finally,

$$d\sigma/dt = V/(RAd)$$

Integrating, one obtains,

$$\sigma(t) = (1/(RAd)) \int_0^t v \quad dt$$

For application purposes,

$$R = 10^6 \text{ ohms}$$

$$d = 33\times 10^{-12} \quad (\text{columb}/m^2)/(N/m^2)$$

$$A = 1.88 \times 10^{-4} \quad m^2$$

Thus, one obtains

$$\sigma(t) = 1.6319 \times 10^8 \int_0^t v\ dt \quad (\text{Pascal})$$

APPENDIX A2 Material Specification of the Test Composite Plate

Manufacturer:	Fiberite Corporation
Fiber:	D-40-700 Graphite
Resin:	Fiberite 974 epoxy
Number of Plies:	Eight
Thickness/ply	0.00537 inch
total thickness	0.038 inch
Density	0.055 lb-mass/cubic inch
Fiber volume	55%

Major Poisson's ratio is 0.25

E11c = Longitudinal compressive modulus = 21.4 Msi

E11t = longitudinal tensile modulus = 24.6 Msi

E22t = transverse tensile modulus = 1.47 Msi

E45t = tensile modulus 45 degree = 3.22 Msi

FLt = Longitudinal ultimate tensile strength = 380 ksi

FLc = longitudinal ultimate compressive strength = 150 ksi

FTt = Transverse ultimate tensile strength = 8.70 ksi

F45t = Ultimate Tensile Strength 45 degree = 27.8 ksi

G12 = shear modulus = 0.67 Msi

FIGURE CAPTIONS

Figure 1 Experimental setup for drop ball stress waves measurement

Figure 2 Schematic diagram of the imbedded sensors and the clamped circular plates and typical stress versus time as sensed by each of the imbedded sensors

Figure 3 Stress parameter versus time for different ball diameters

Figure 4 The two waveforms as sensed by the two imbedded sensors recorded on the high speed digital waveform recording oscilloscope

Figure 5a An enlarged photograph from the oscilloscope showing the delay in the starting times from the two sensors

Figure 5b Waveforms from the PVDF sensors show repeatability of of the test in eight consecutive drop ball impacts

Figure 6 Front part of the plot of two waves generated by the drop of a one millimeter diameter steel ball.

Figure 7 Amplitude decay of the stress wave as it propagates through the three consecutive snsors located at seven ply intervals

Figure 8 Stress versus time for the three imbedded sensors

Figure 9 the wave sensed by the first sensor is replotted to show superposition of compressive and tensile waves

Figure 10 Superposition of wave as sensed by all three sensors

Figure A1 PVDF sensor measurement principle and typical votage vs time and stress versus time curves

Figure A2 Schematic piezopolymer PVDF film

Figure 1 Experimental setup for drop ball stress waves measurement

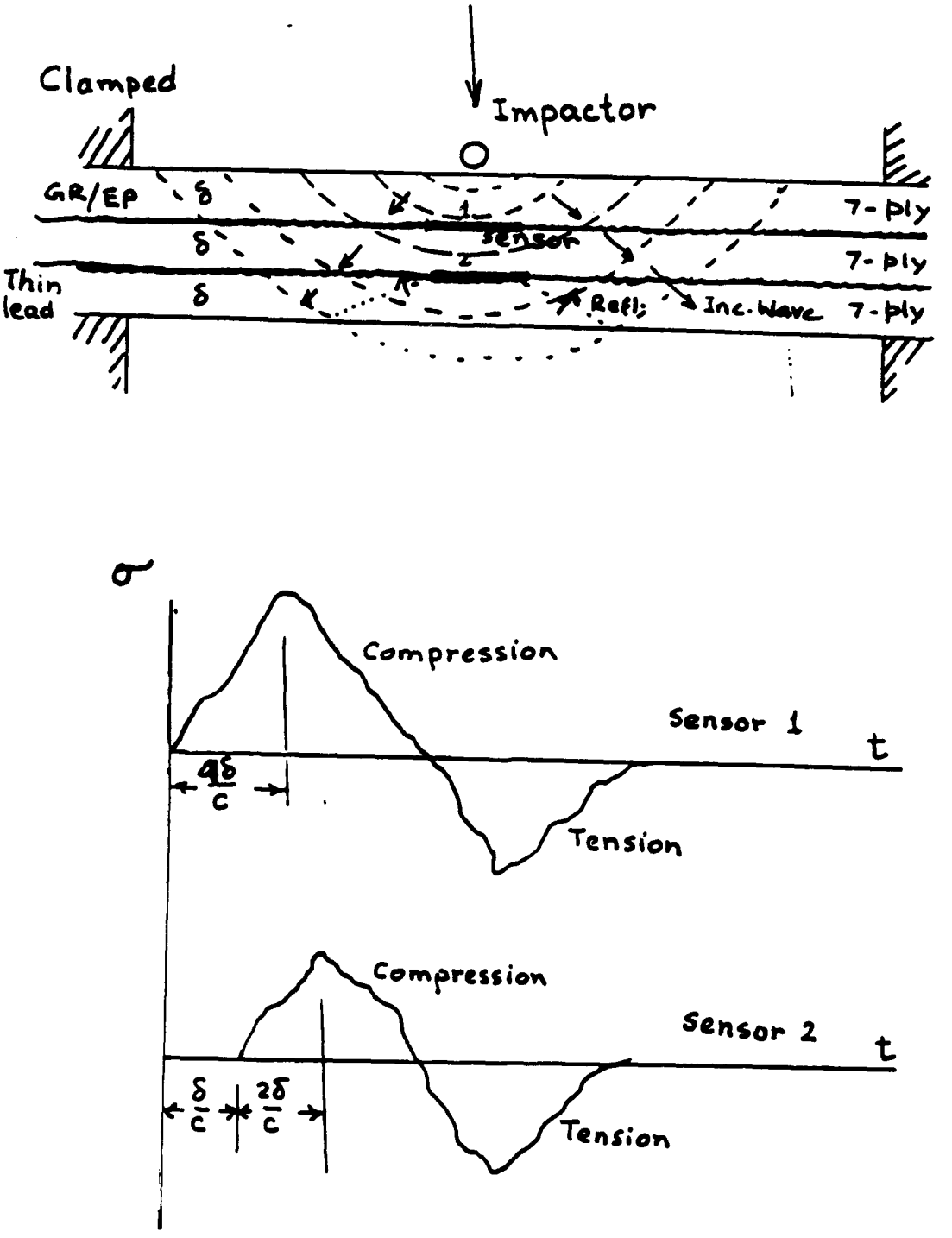


Figure 2 Schematic diagram of the imbedded sensors and the clamped circular plates and typical stress versus time as sensed by each of the imbedded sensors

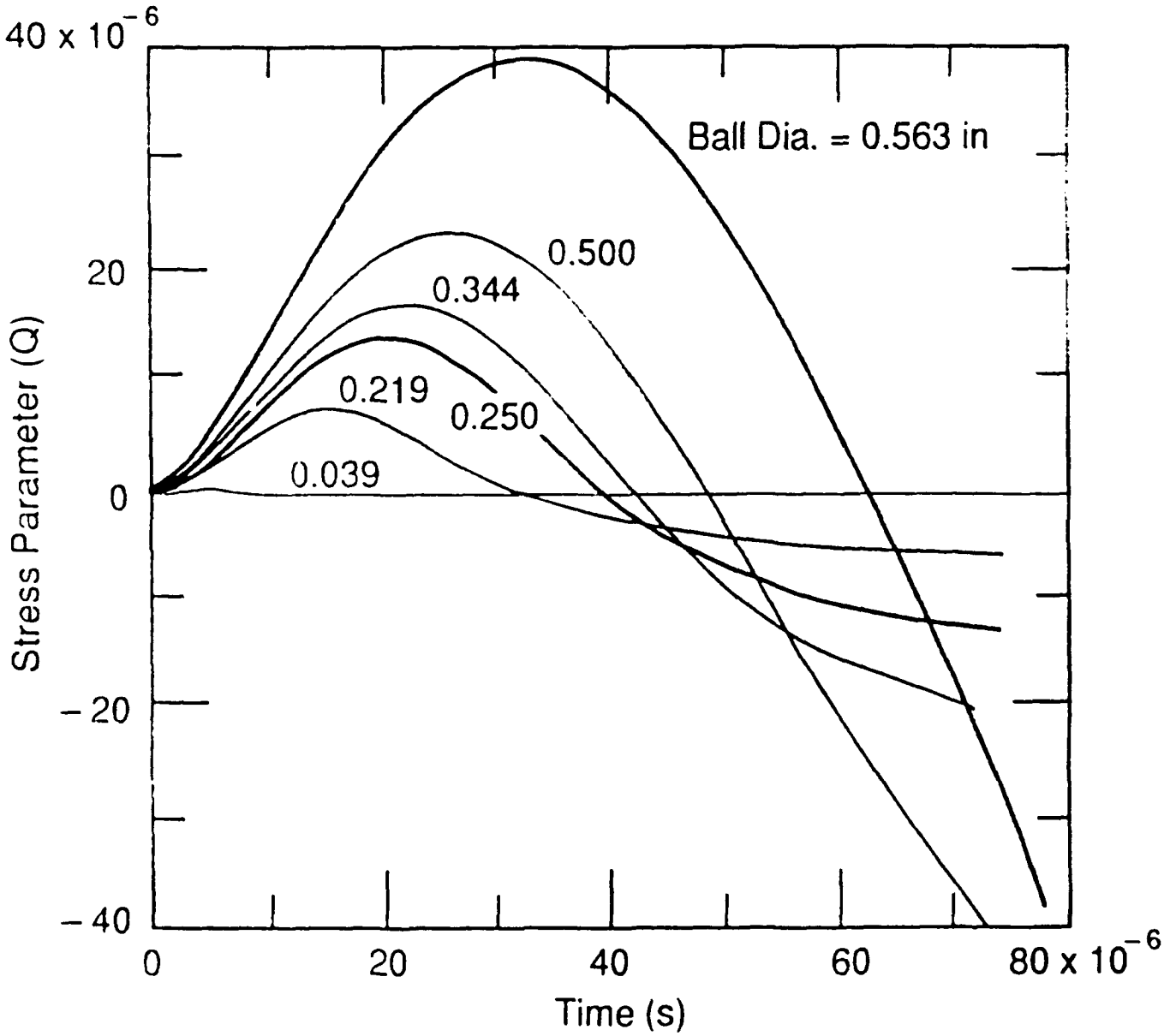


Figure 3 Stress parameter versus time for different ball diameters

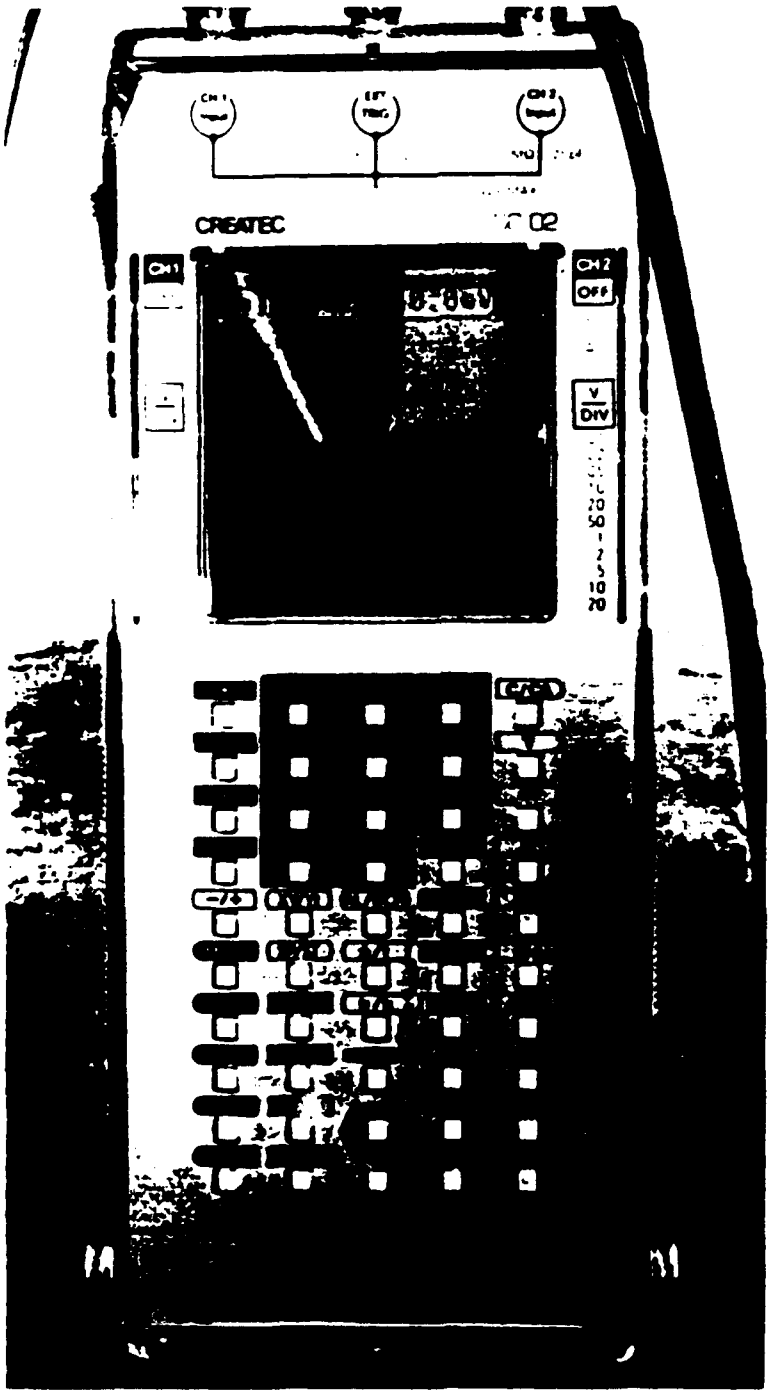


Figure 4 The two waveforms as sensed by the two imbedded sensors recorded on the high speed digital waveform recording oscilloscope

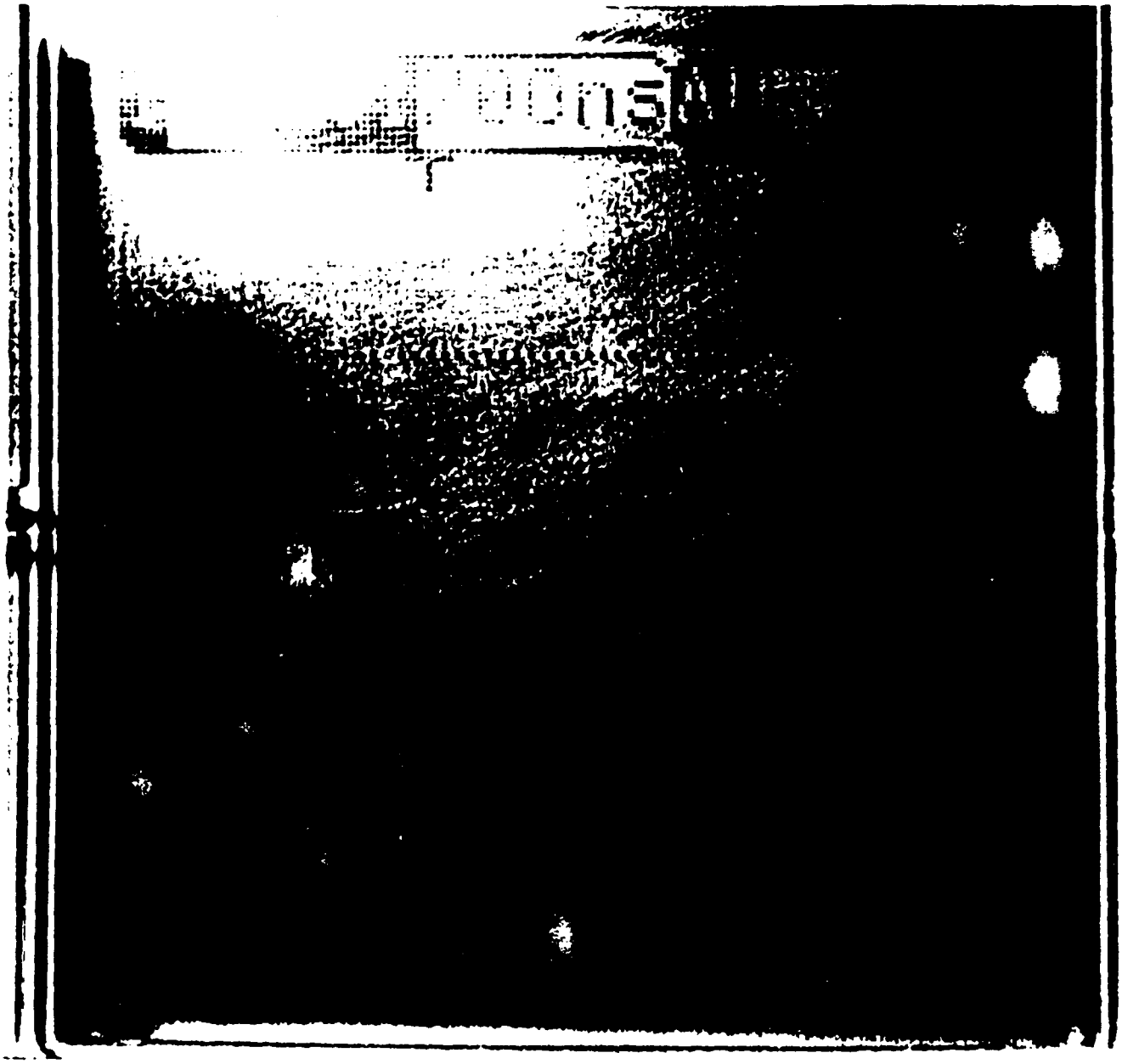
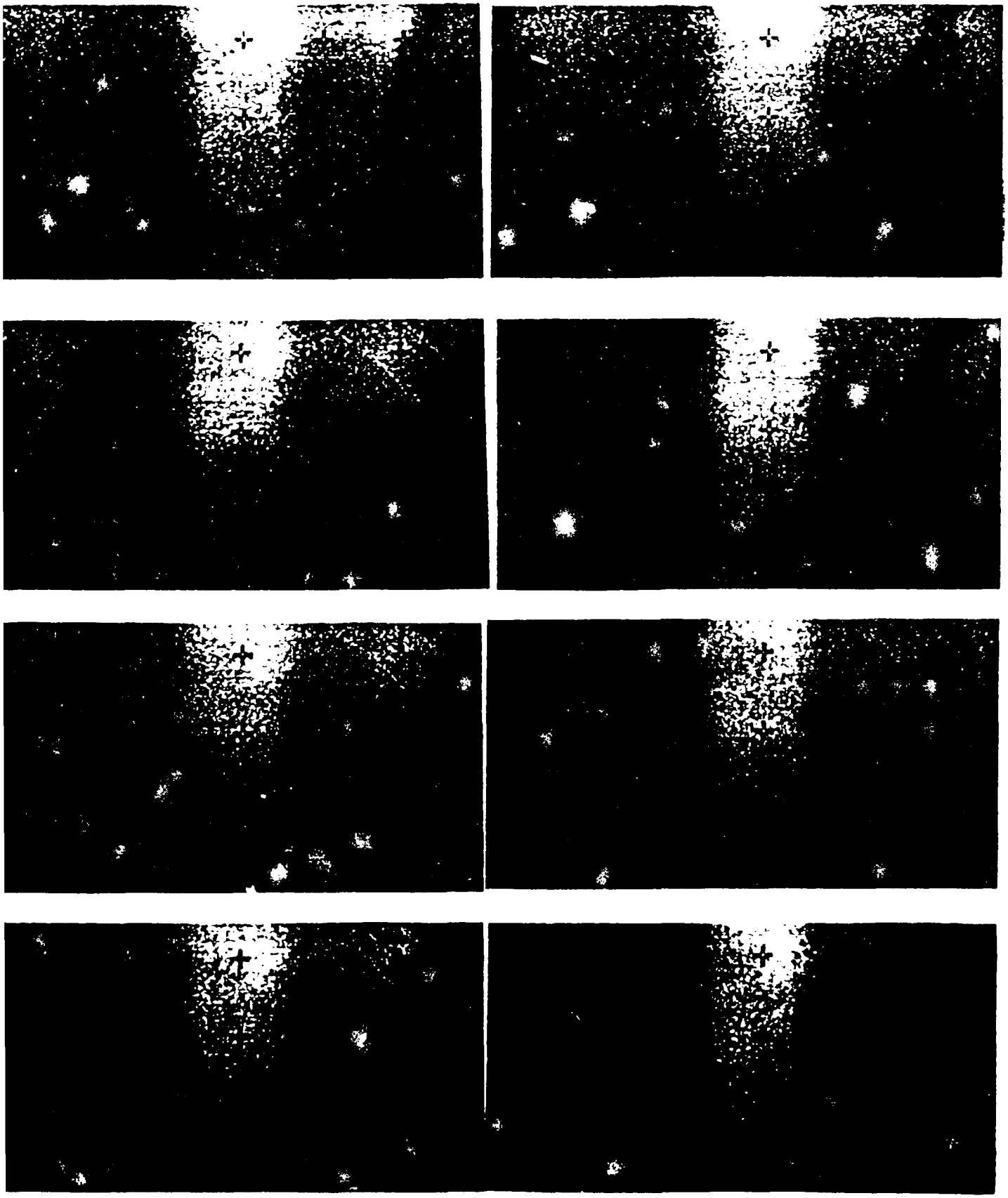


Figure 5a An enlarged photograph from the oscilloscope showing the delay in the starting times from the two sensors



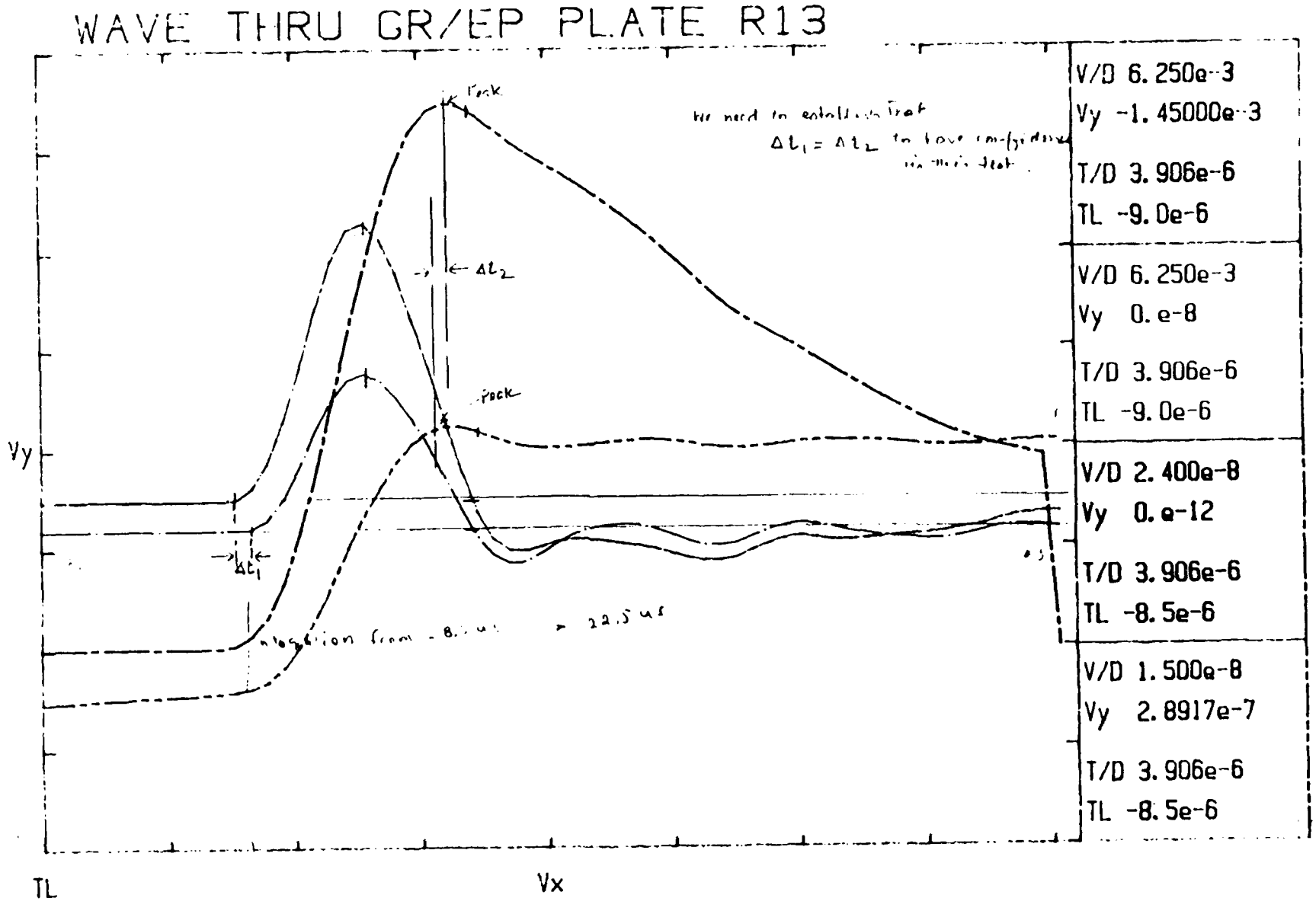


Figure 6 Front part of the plot of two waves generated by the drop of a one millimeter diameter steel ball.

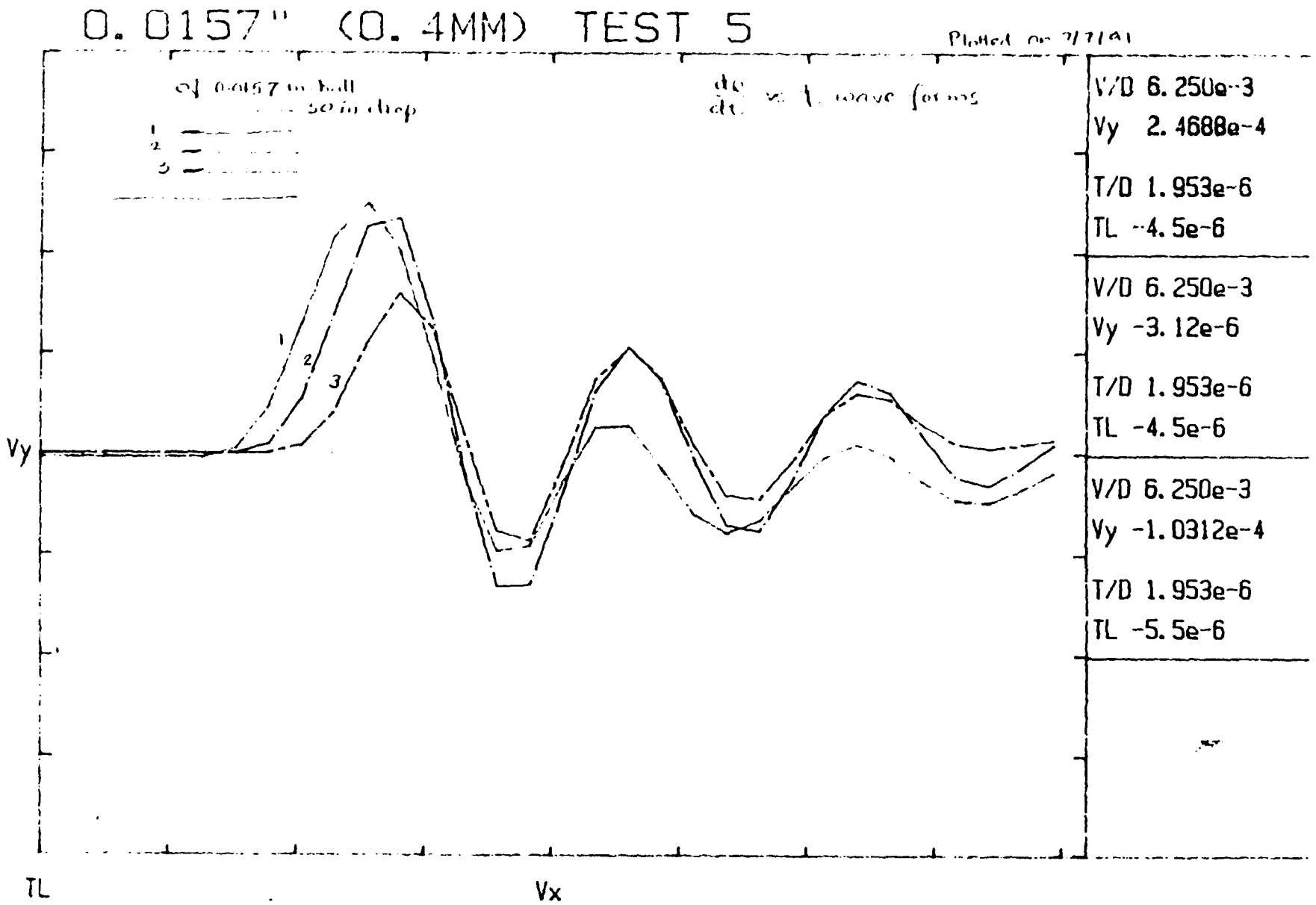


Figure 7 Amplitude decay of the stress wave as it propagates through the three consecutive snsors located at seven ply intervals

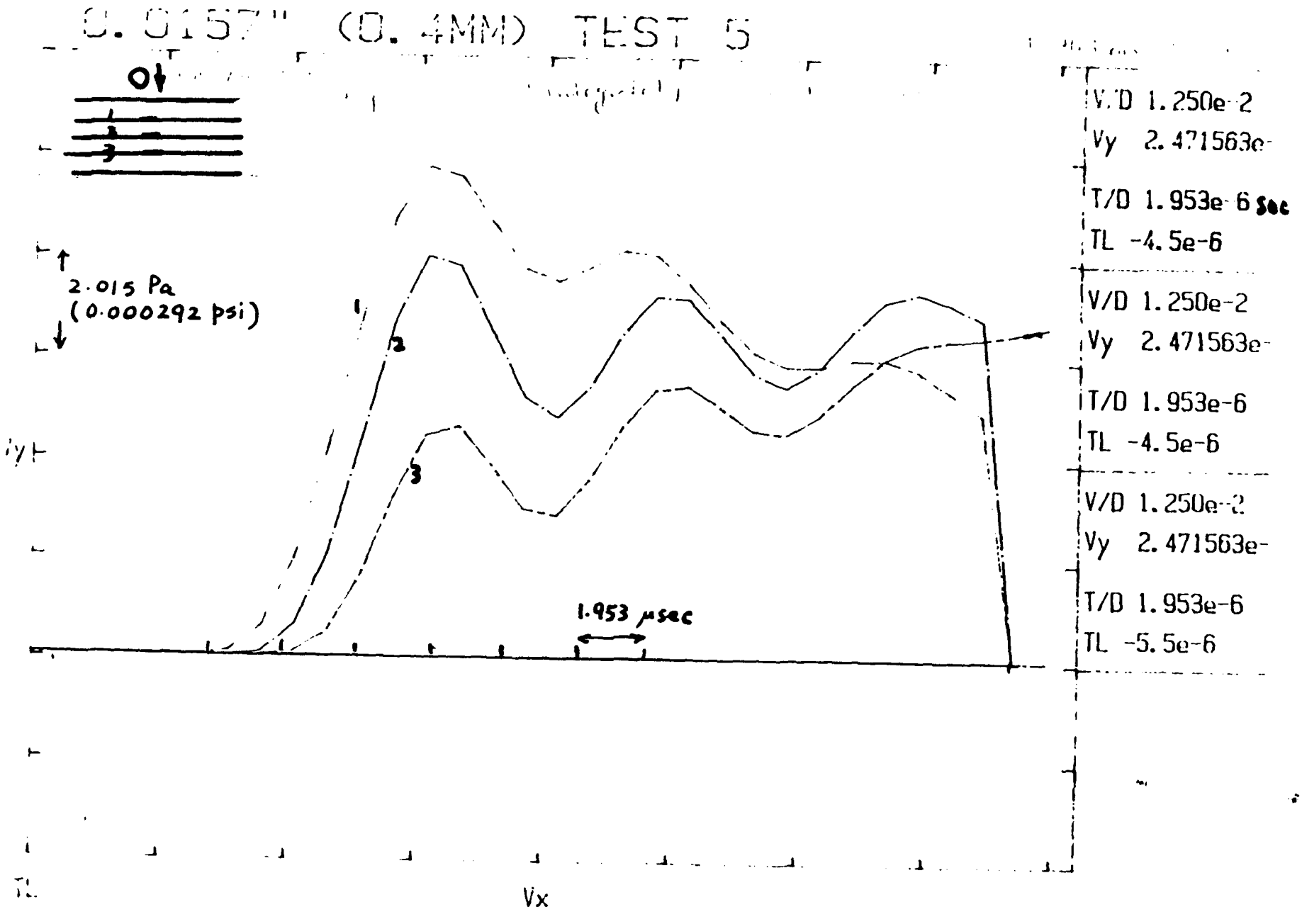


Figure 8 Stress versus time for the three imbedded sensors

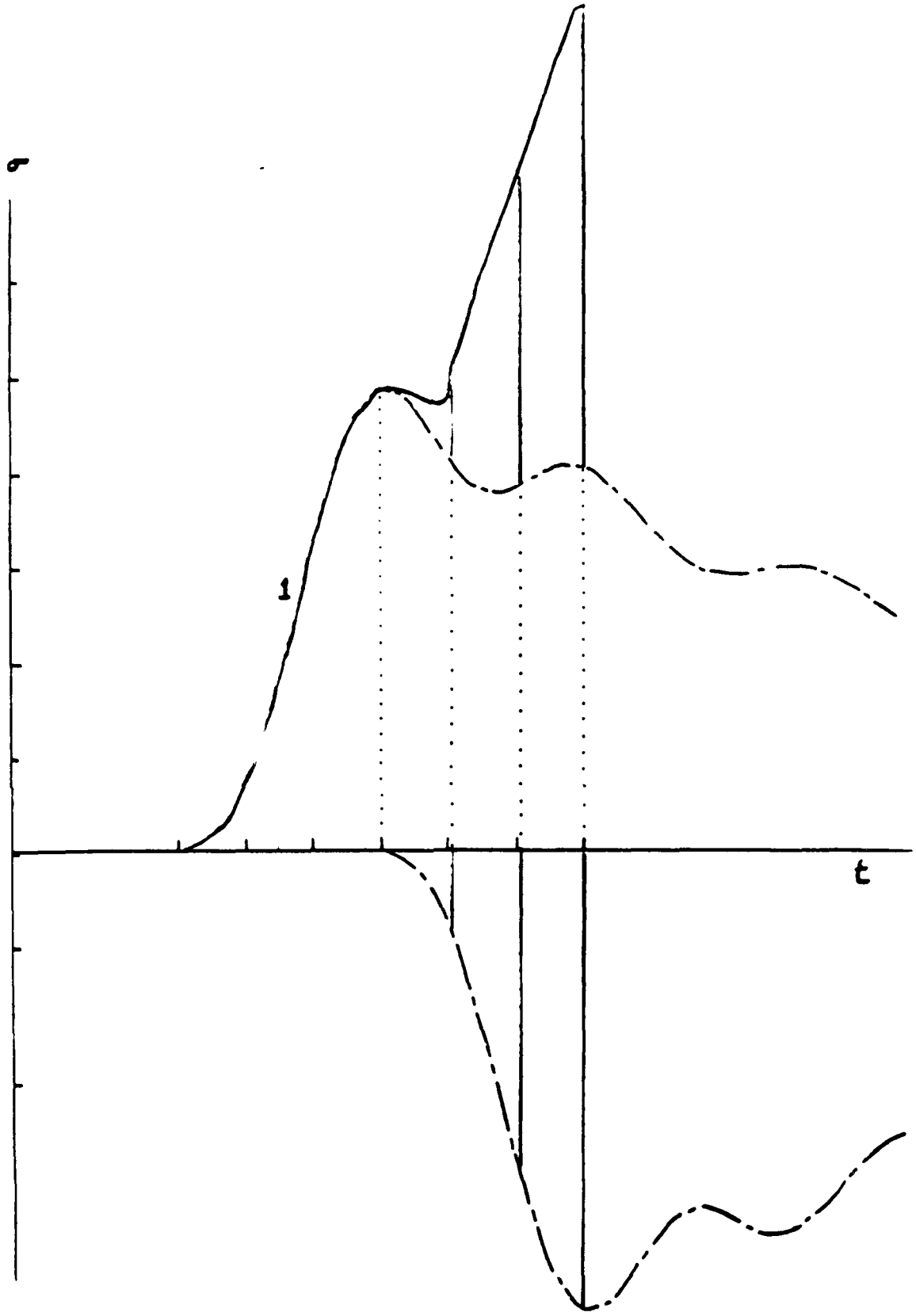


Figure 9 The wave sensed by the first sensor is replotted to show superposition of compressive and tensile waves

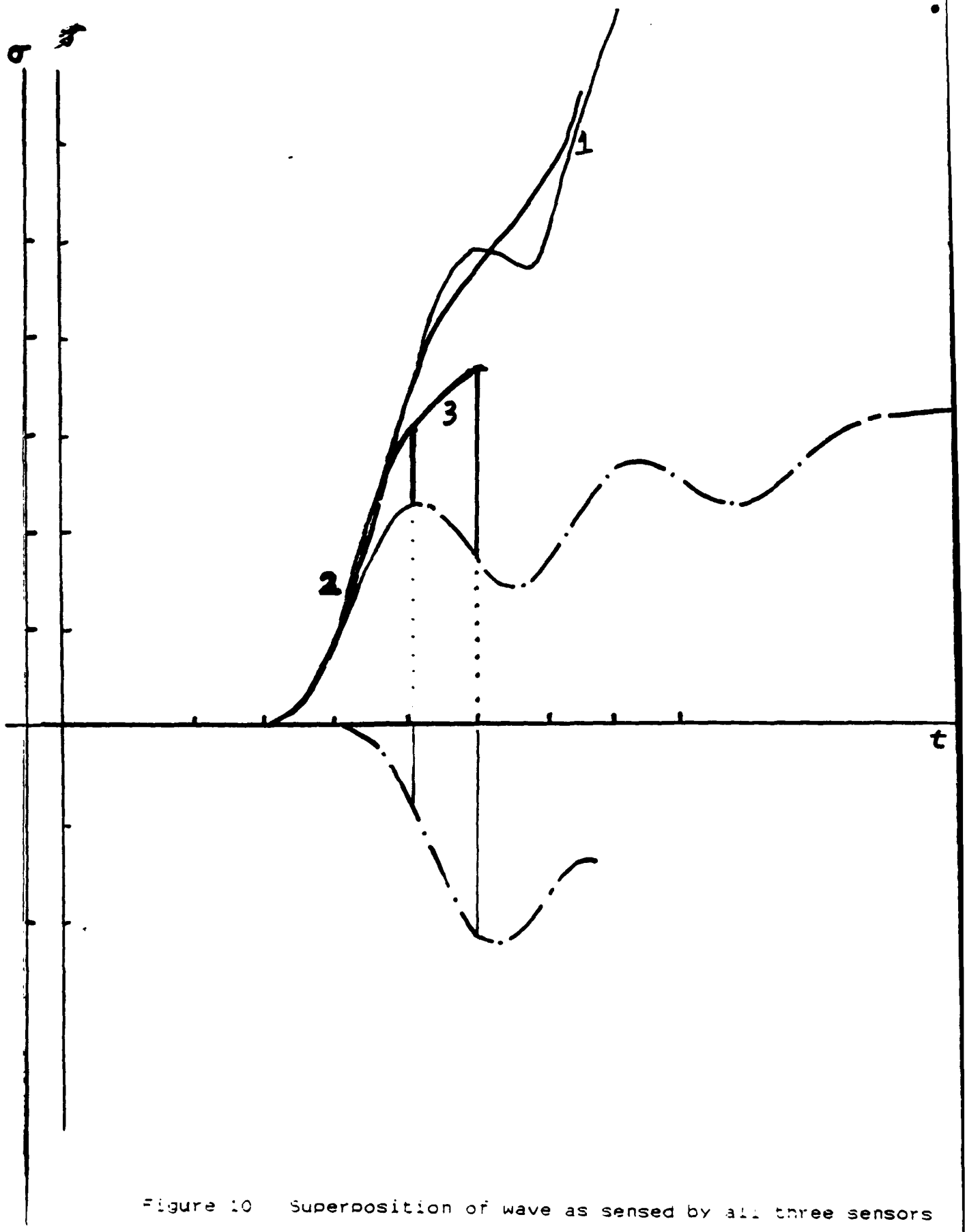


Figure 10 Superposition of wave as sensed by all three sensors

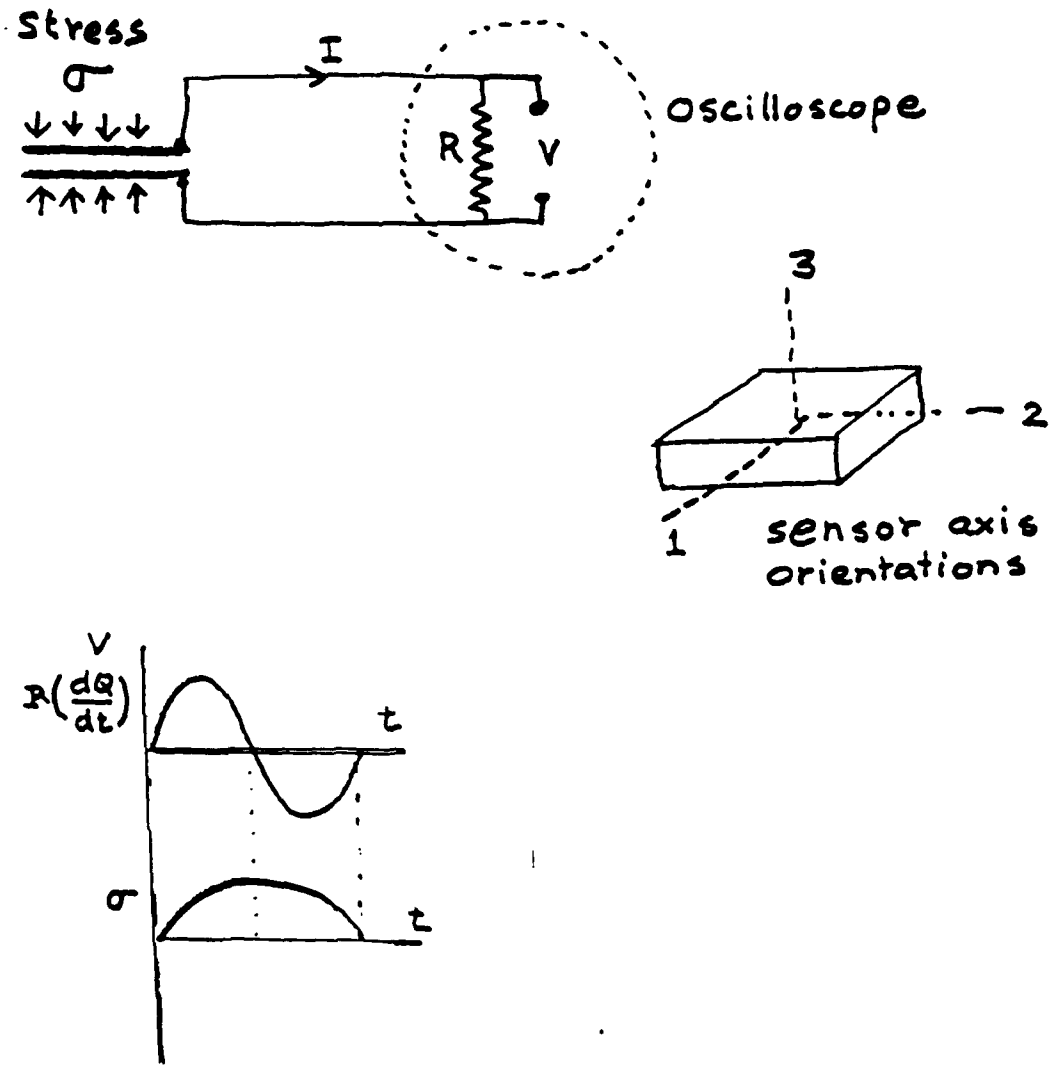


Figure A1 PVDF sensor measurement principle and typical votage vs time and stress versus time curves

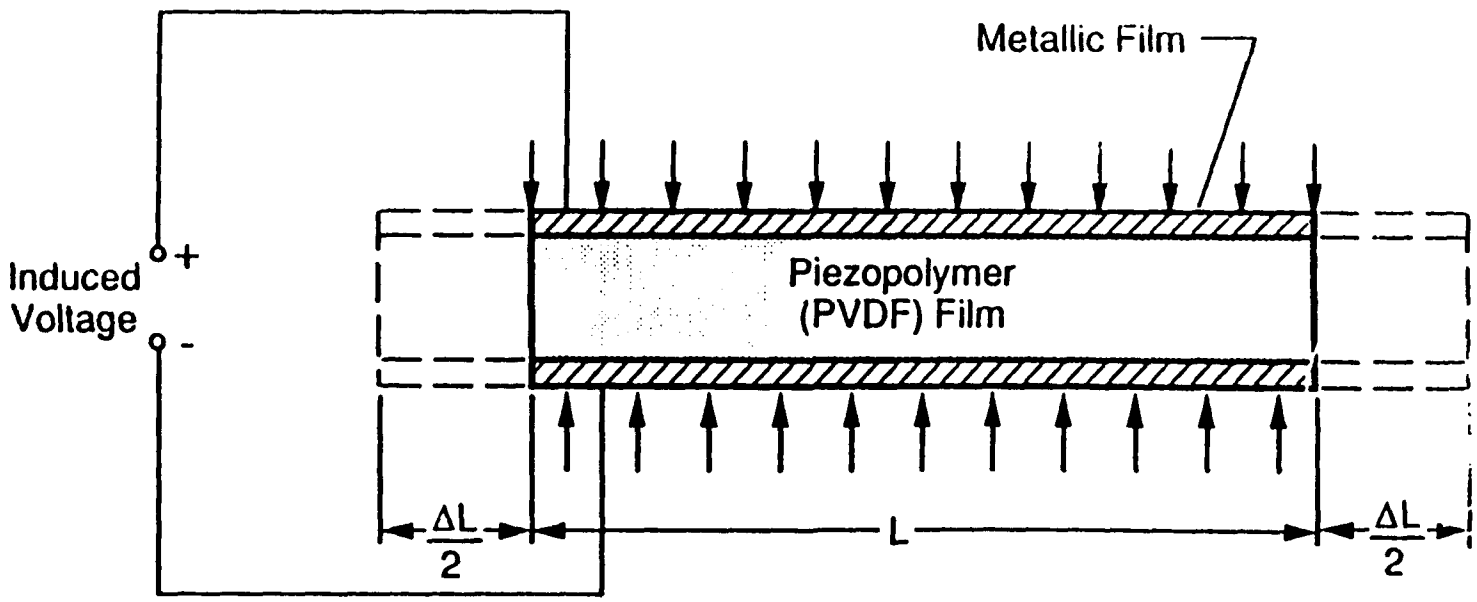


Figure A2 Schematic piezopolymer PVDF film

ACCELERATED FATIGUE TEST PROCEDURE FOR THE STRUCTURAL POLYCARBONATE COMPONENT OF THE F-16 CANOPY COMPOSITE MATERIAL.

YULIAN B. KIN
PROFESSOR OF MECHANICAL ENGINEERING
PURDUE UNIVERSITY CALUMET
SCHOOL OF PROFESSIONAL STUDIES
ENGINEERING DEPARTMENT

THE PROJECT IS SPONSORED BY THE AIR FORCE OFFICE OF SCIENTIFIC RESEARCH AND CONDUCTED BY THE UNIVERSAL ENERGY SYSTEMS, INC.
CONTRACT NO. 9620-88-C-0053/SB5881-0378
P.O. NO. S-210-11MG-005

DECEMBER, 1991

ABSTRACT

The existing long-term fatigue test procedure requires the breaking of 20 to 30 identically-prepared specimens and one month to complete. Thus, manufacturers do not perform a conventional fatigue test in spite of its obvious utility. Therefore, there is a definite need for an accelerated fatigue test which can be completed in approximately one day. The accelerated test procedure proposed in this project was developed on the basis of data obtained during the conventional fatigue test performed under UES mini grant S-210-9MG-038 in 1989, and as a continuation of research started by the principal investigator at the Flight Dynamics Directorate, Wright-Patterson Air-Force Base, Summer 1990. The accelerated test proposed will permit to control a stability of the manufacturing process and enable preliminary estimates of the design changes.

The additional long-term tests were run to check a compatibility of the new and previous test results, gain more statistics, and increase reliability of the basis information.

The fatigue experimental investigation was conducted with polycarbonate specimens of different thicknesses both with and without stress concentrators. The influence of the frequency on the fatigue life of these specimens was also investigated.

ACKNOWLEDGEMENTS

The project was developed as a continuation of research started by the project principal investigator during his Summer 1990 tenure at the Flight Dynamics Directorate, Wright Laboratories, Wright-Patterson Air Force Base.

Sponsorship by the AIR FORCE OFFICE OF SCIENTIFIC RESEARCH, BOLLING AFB, DC and financial support from Universal Energy Systems, Inc. (Mini Grant Award S-210-11Mg-005) is gratefully acknowledged.

Particular thanks to Aerospace Engineer Lorene V. Garrett, Aircrew Protection Branch Supervisor Robert E. McCarty, and Dr. Arnold Mayer for their constant interest and technical support during different phases of this project.

Purdue University Calumet provided matching financial and equipment support to make this project possible.

LIST OF FIGURES AND TABLES

Figure 1. Loading program for the accelerated test. Three fatigue curves A,B, and C are received from a long-term fatigue test and correspond, for example, 5%, 50%, and 95% probability of failure.

Figure 2. A diagram for graphical determination of fatigue strength.

Figure 3. MTS flexure fatigue fixture.

Figure 4. Loading diagram.

Figure 5. Specimen geometry.

Figure 6. S-N diagram plotted from the results of fatigue test for the solid polycarbonate specimens.

Figure 7. S-N diagram plotted from the results of fatigue test for the polycarbonate specimens with stress concentrators.

Figure 8. S-N diagram for polycarbonate specimens with stress concentrators. 1991 year fatigue test. A, B, C curves plotted to calculate relative lives during treatment of the accelerated test results.

Figure 9. Program and result of accelerated fatigue test. Specimen #1 with stress concentration. Specimen dimensions in Figure 5.

Figure 10. Program and result of accelerated fatigue test. Specimen #2 with stress concentration. Specimen dimensions in Figure 5.

Figure 11. Graphical representation of the accelerated fatigue test. Specimen #1. The test program is in Figure 9.

Figure 12. Comparison of fatigue lives of the specimens with different thickness. Variable bending. Specimen with stress concentration. frequency 5 Hz. Five specimens tested in each group.

Table 1. Testing regimes (Long-term test; variable bending)

Table 2. Accelerated test results.

Table 3. Accelerated test results treatment.

Table 4. Fatigue lives of specimens tested at different frequencies. Specimens with stress concentration. Amplitude load is 240 Lb.

INTRODUCTION

It appears (based on preliminary tests) that the fatigue properties of polycarbonate sheets vary significantly from sheet to sheet, and probably within the same sheet. Therefore, it is very important to have a mechanism which permits (quick and continuous) quality control of polycarbonate sheets and detects deviations in the manufacturing process; in other words, a mechanism to control the stability of a manufacturing process. The accelerated fatigue tests can be utilized for such control, and the test procedure developed in this project can be useful for this purpose. The accelerated fatigue tests would also be useful for preliminary evaluations of new designs. For example, the Flight Dynamics Directorate is developing frameless canopy design and the accelerated fatigue test used in this project can be recommended for a preliminary estimate of fatigue resistance of the frameless canopy polycarbonate.

The accelerated fatigue test procedure proposed in this project was partially developed on the basis of data obtained during the long-term fatigue test performed under UES mini grant S-210-9MG-038 in 1989, and as a continuation of research started by the principal investigator at the Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Summer 1990.

OBJECTIVES

The major objectives of the study were:

1. Development of detailed procedure for the accelerated fatigue test for coupons cut from a structural polycarbonate sheet of F-16 canopy laminate.

2. Conducting additional long-term fatigue tests using the procedure given in reference 1. Additional fatigue tests are necessary to get reliable parameters for the accelerated fatigue tests.

3. Conducting accelerated fatigue tests based on the procedure developed and using statistics gained from long-term fatigue tests.

The objectives of the initiation investigations for the possible future detailed research were:

1. Determination of a dependency of fatigue resistance of polycarbonate sheet on thickness of the sheet.

2. Determination of fatigue lives of specimens tested at different frequencies.

LOCATY'S PROCEDURE FOR ACCELERATED FATIGUE TEST

The theory and description of the Locaty's accelerated fatigue test had been given in [1] and [2] and repeats here in concise form for the convenience.

The method is based on the concept of cumulative fatigue damage or Palmgren-Minor rule considering $\sum\left(\frac{n_i}{N_i}\right) = 1\text{-}5$, where n_i is the number of cycles which specimen worked in the specified test regime, and N_i is the number of cycles which specimen could potentially work in accordance with the fatigue curve received from the results of the long-term fatigue tests of the same type of the specimens.

The loading program and the treatment of results are presented in Figures 1 and 2. Figure 1 shows three fatigue curves (for example, 5%, 50%, and 90% probability of failure) received from a long-term fatigue test. During an accelerated test the specimen works, for example, for 50,000 cycles at the first load level, then the load is increased and the specimen is again tested for 50,000 cycles. The procedure is repeated until the instant when the specimen fails. Then using Figure 1, the magnitudes of

$$\sum\left(\frac{n_i}{N_i}\right)_A, \quad \sum\left(\frac{n_i}{N_i}\right)_B, \quad \sum\left(\frac{n_i}{N_i}\right)_C$$

are determined. With these three parameters (or, if necessary, a greater number of points) and knowing the corresponding stresses, we can find the coordinates of the points which result in the curve shown in figure 2. Now, if according to an accepted hypothesis fatigue strength corresponds to a definite value (for example, $\sum\left(\frac{n_i}{N_i}\right) = 1$, it is possible to determine the magnitude of fatigue strength (Figure 2).

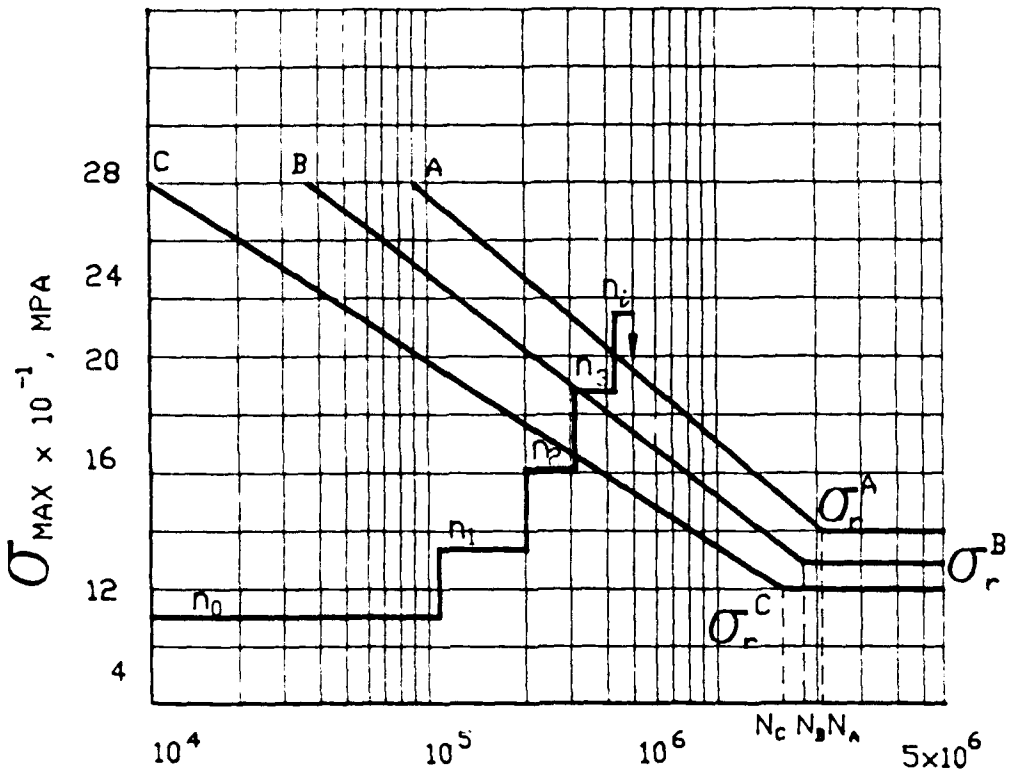


Figure 1. Loading program for the accelerated test. Three fatigue curves A, B, C are received from a long-term fatigue test and correspond, for example, 5%, 50%, and 95% probability of failure.

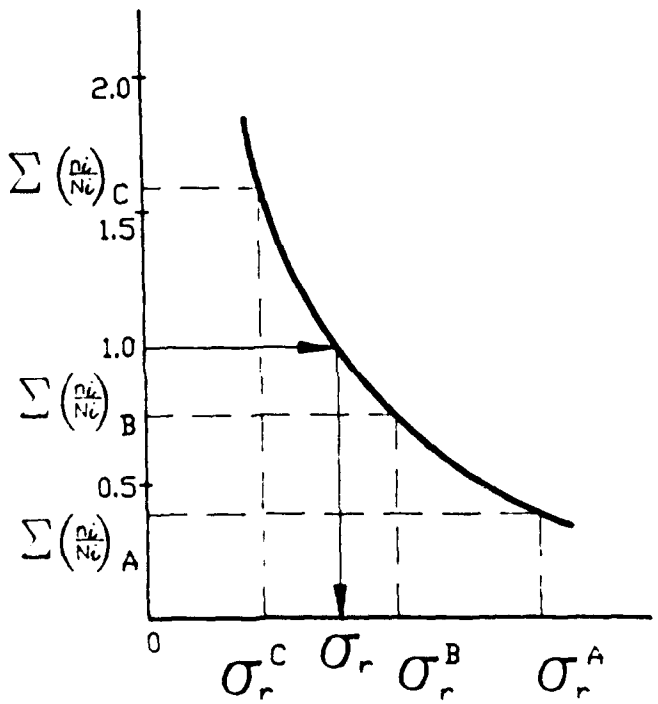


Figure 2. A diagram for graphical determination of fatigue strength.

EQUIPMENT, SPECIMENS, AND PROCEDURE OF LONG-TERM FATIGUE TESTS

The flexure conventional and accelerated fatigue tests were conducted on an MTS machine using the four point MTS flexure systems to provide bending (Figure 3). The loading diagram is shown in Figure 4. The specimen geometry is given in Figure 5. It can be noted that the testing machine, fixtures, specimens with stress concentrators, loading diagram, and load spans were used absolutely the same as described in [1] and [2] to provide a compatibility of the previous and new tests.

The specimen machining procedure, conventional fatigue test procedure, and failure criteria were also similar to described in [1] and [2].

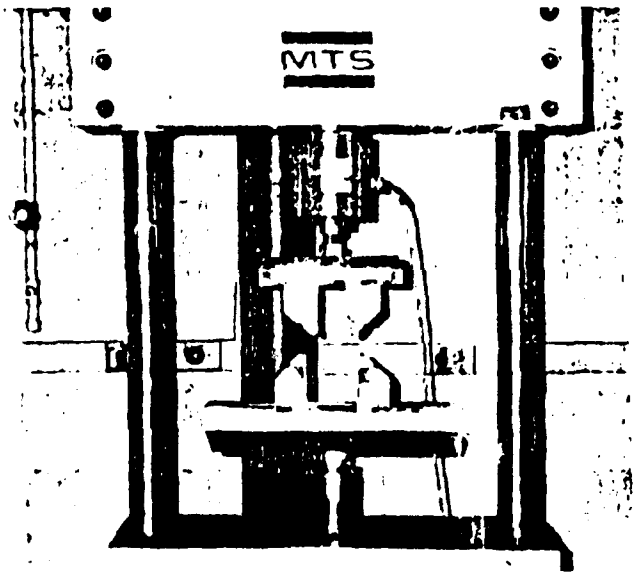
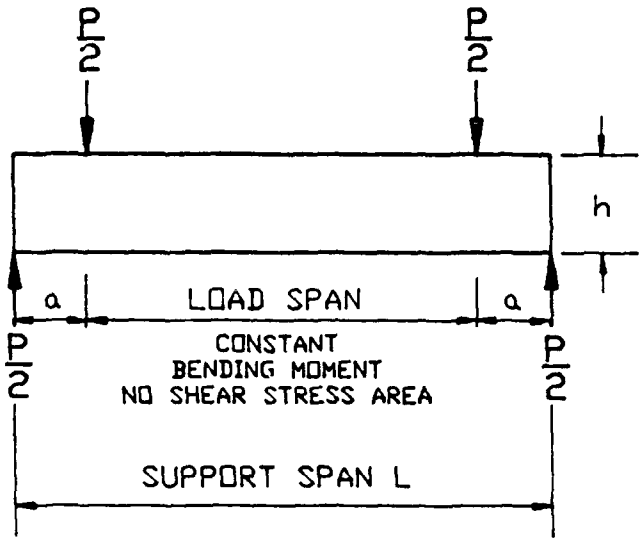


Figure 3. MTS flexure fatigue fixture



MAX DEFLECTION:

$$Y_{MAX} = \frac{Pa}{48EI}(4a^2 - 3L^2)$$

MAX STRESS:

$$\sigma_{MAX} = \frac{3P(a)}{bh^2}$$

Figure 4. Loading diagram

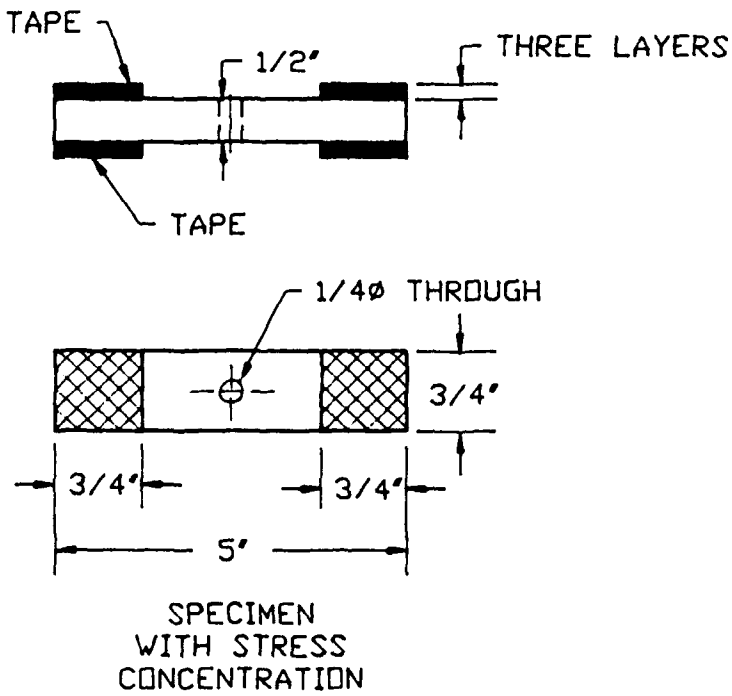


Figure 5. Specimen geometry

LONG-TERM AND ACCELERATED FATIGUE TESTS

The long-term fatigue test regimes are given in Table 1, and the programs and parameters of the accelerated tests are given in Figures 8,9 and 10.

Table 1
Testing Regimes (Long- term test; variable bending).

Amplitude load, Lb.		Amplitide stress, psi	Min. stress / Max. stress	Frequency, Hz
Plate wth a hole	Solid plate			
180	240	1440	0.2	12
	360	2160	0.2	8
360	480	2880	0.2	6
	600	3600	0.2	5
540	720	4320	0.2	4
750	990	5760	0.2	3

The results of the long-term fatigue tests from [1] and [2] are shown in Figures 6 and 7 and repeated in this report for the comparison purposes. The results long-term fatigue tests completed in 1991 are shown in Figure 8. From the results gained it can be assumed that the 1989 material had greater fatigue strength than the polycarbonate sheet received in 1990 and tested in 1990 and 1991 years. But the scatter is significantly less in the last tests. The possible difference in strength of two polycarbonate sheets of the same type shows one more time that the quick control of fatigue characteristics is very useful to verify the initial product quality.

The results and programs of the accelerated tests are given on loading diagrams in Figures 8, 9, 10 and in Table 2. The treatment of the test data is shown in Table 3. The testing lives

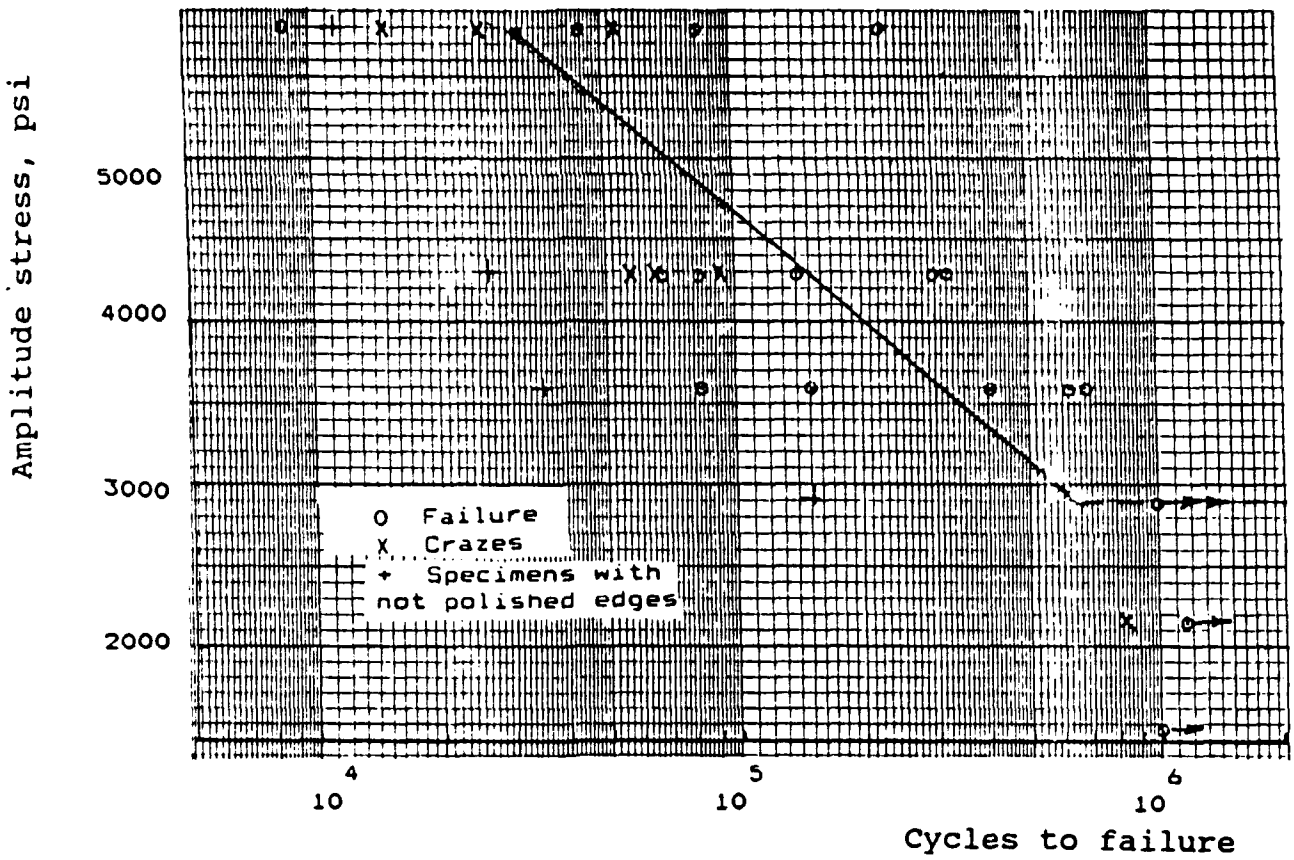


FIGURE 6. S-N Diagram plotted from the results of fatigue test for the solid polycarbonate specimens.

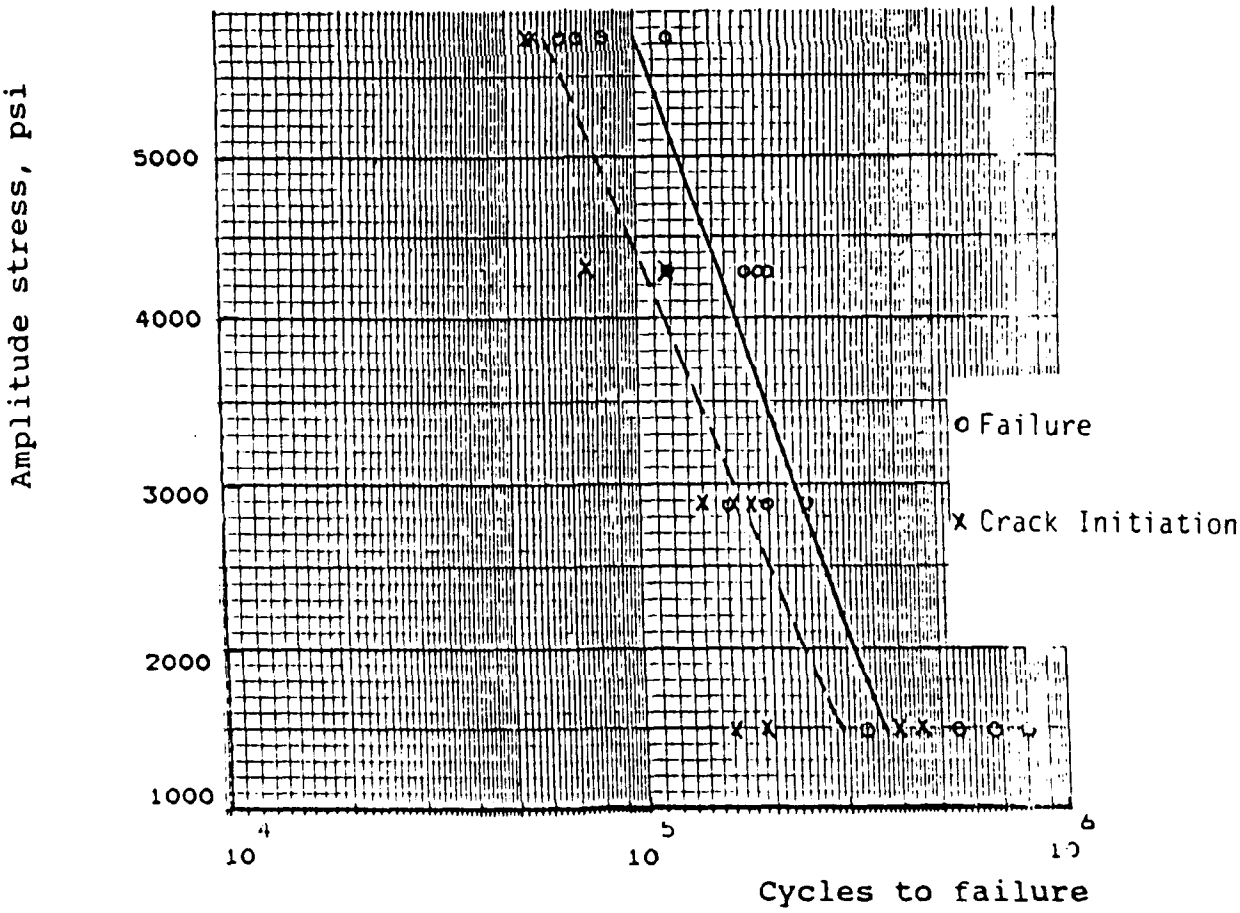


FIGURE 7. S-N Diagram plotted from the results of fatigue test for the polycarbonate specimens with stress concentrators.

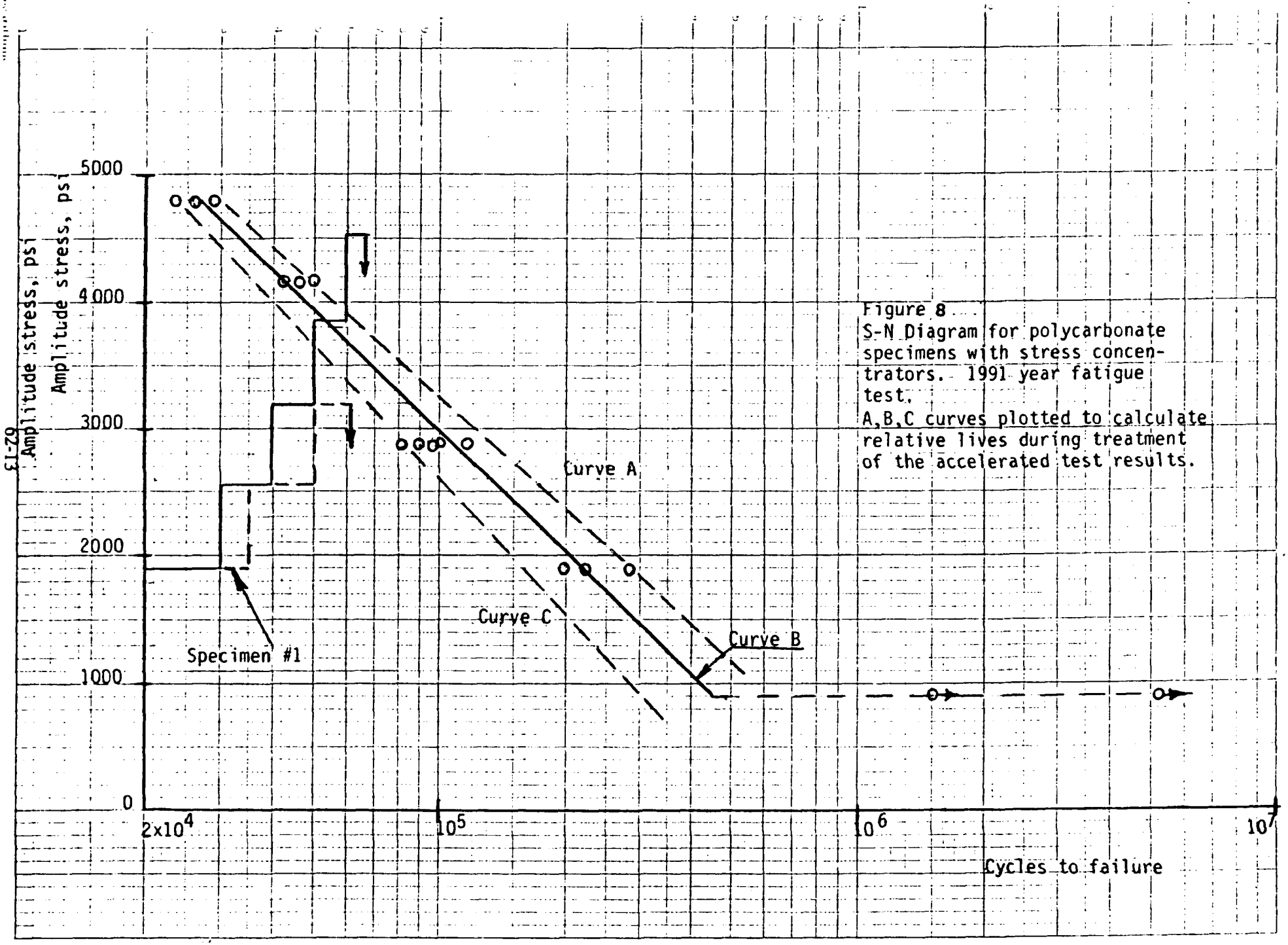


Figure 8
S-N Diagram for polycarbonate specimens with stress concentrators. 1991 year fatigue test.
A,B,C curves plotted to calculate relative lives during treatment of the accelerated test results.

TABLE 2. Accelerated test results.

Specimen number	Max. load increment, Lb	Number of cycles at one load level	Number of steps until failure	Number of cycles until failure
1	200	15000	4	56000
2	200	15000	5	74200
3	200	15000	4	50500
4	200	15000	3	53900

TABLE 3. Accelerated test result treatment.

Amplitude stress, psi	Curve A		Curve B		Curve C	
	Ni, cycles	n /Ni	Ni, cycles	n /Ni	Ni, cycles	n /Ni
1280	660000	0.023	550000	0.027	280000	0.054
1920	480000	0.031	320000	0.047	240000	0.054
2560	250000	0.060	135000	0.111	105000	0.143
3200	110000	0.100	85000	0.129	68000	0.162
		0.214		0.314		0.422

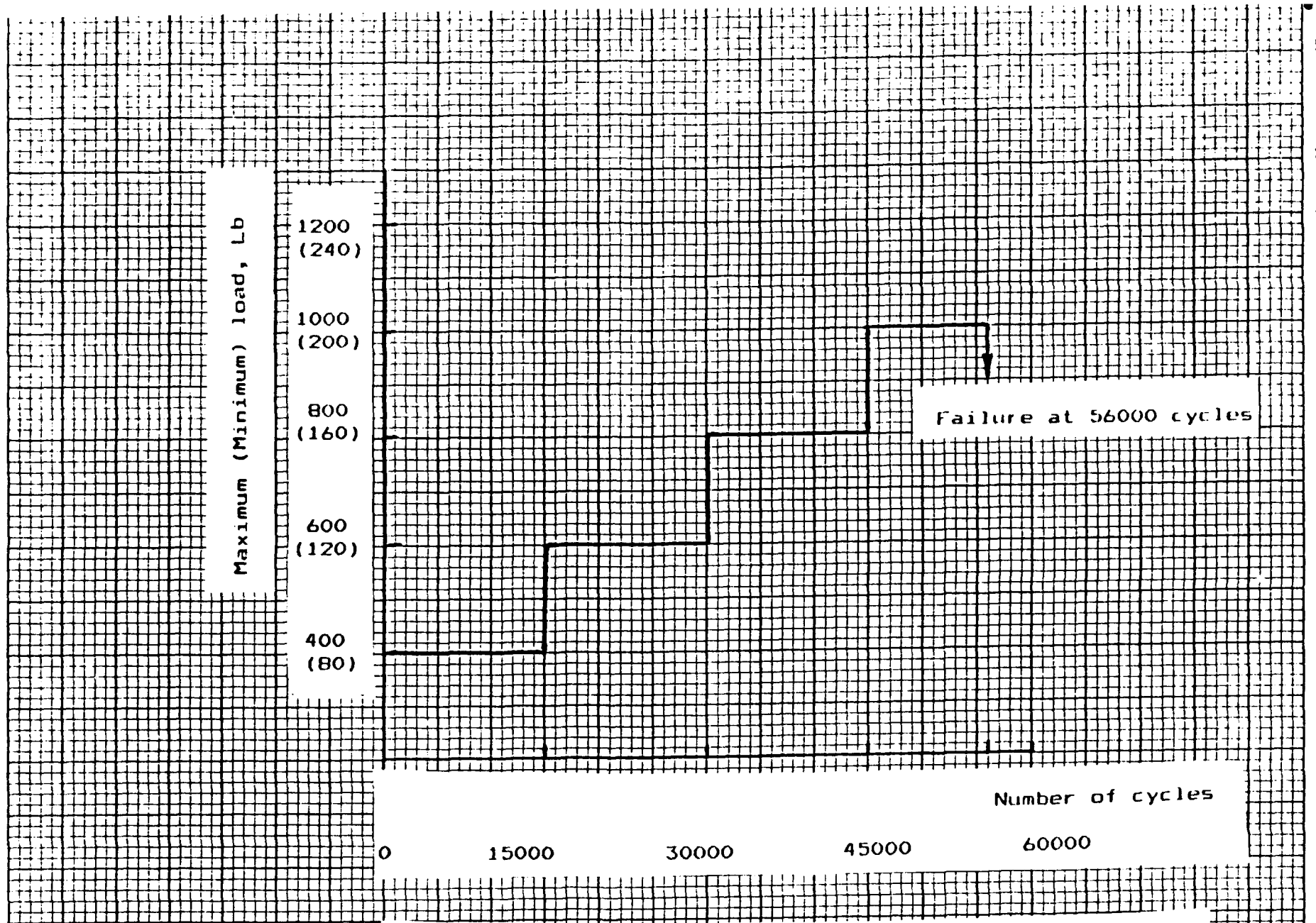


Figure 9. Program and result of accelerated fatigue test. Specimen #1 with stress concentration. Specimen dimensions in figure 2.

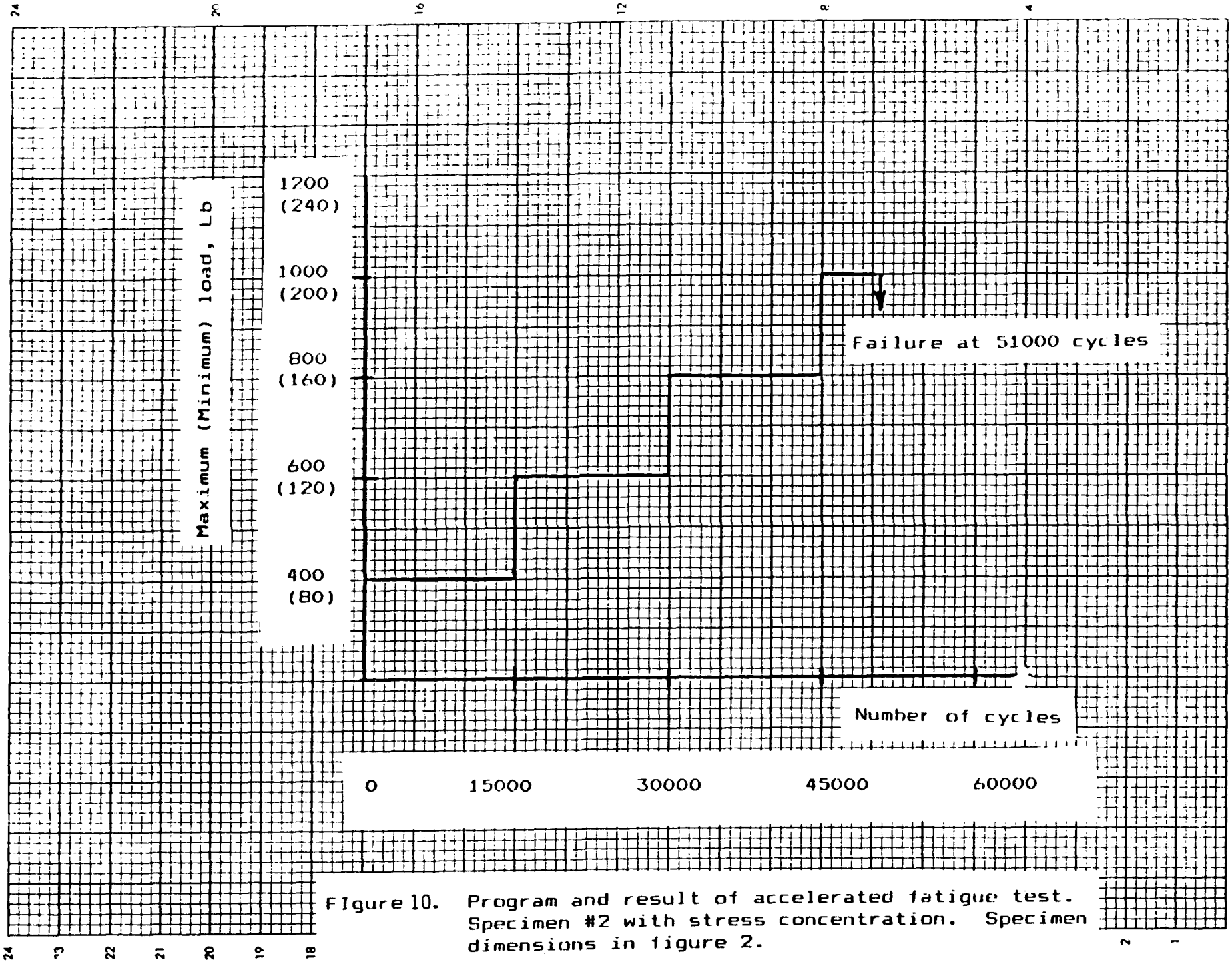


Figure 10. Program and result of accelerated fatigue test. Specimen #2 with stress concentration. Specimen dimensions in figure 2.

n_i are taken from accelerated test program which is given, for example, in Figure 9, and expected lives N_i for the curves A, B, and C (90%, 5%, and 95% probability of survival) are taken from the Figure 8. The repeatability of the results is quite reasonable and, therefore, the sum of relative lives which is determined experimentally can probably be recommended as a basic parameter to confirm whether the specimens tested during the control procedure belong to the entire population. The time of accelerated tests was never more than 6 hours. It is even probably possible to decrease the testing time trying different combinations of the load increment and number of testing cycles at each load level, but it was beyond the scope of this initiation research project.

The failure damage and failure mechanism during long-term and accelerated fatigue tests completed in 1991 year were similar to those which were described in [1] and [2]. The cracks always started at the bottom tensile zones of the specimens. In all specimens with stress concentrators the cracks propagated from the hole edges toward the specimens sides. Usually a minute crack spot preceded the crack formation and propagated ahead of the crack tip during the entire damage process.

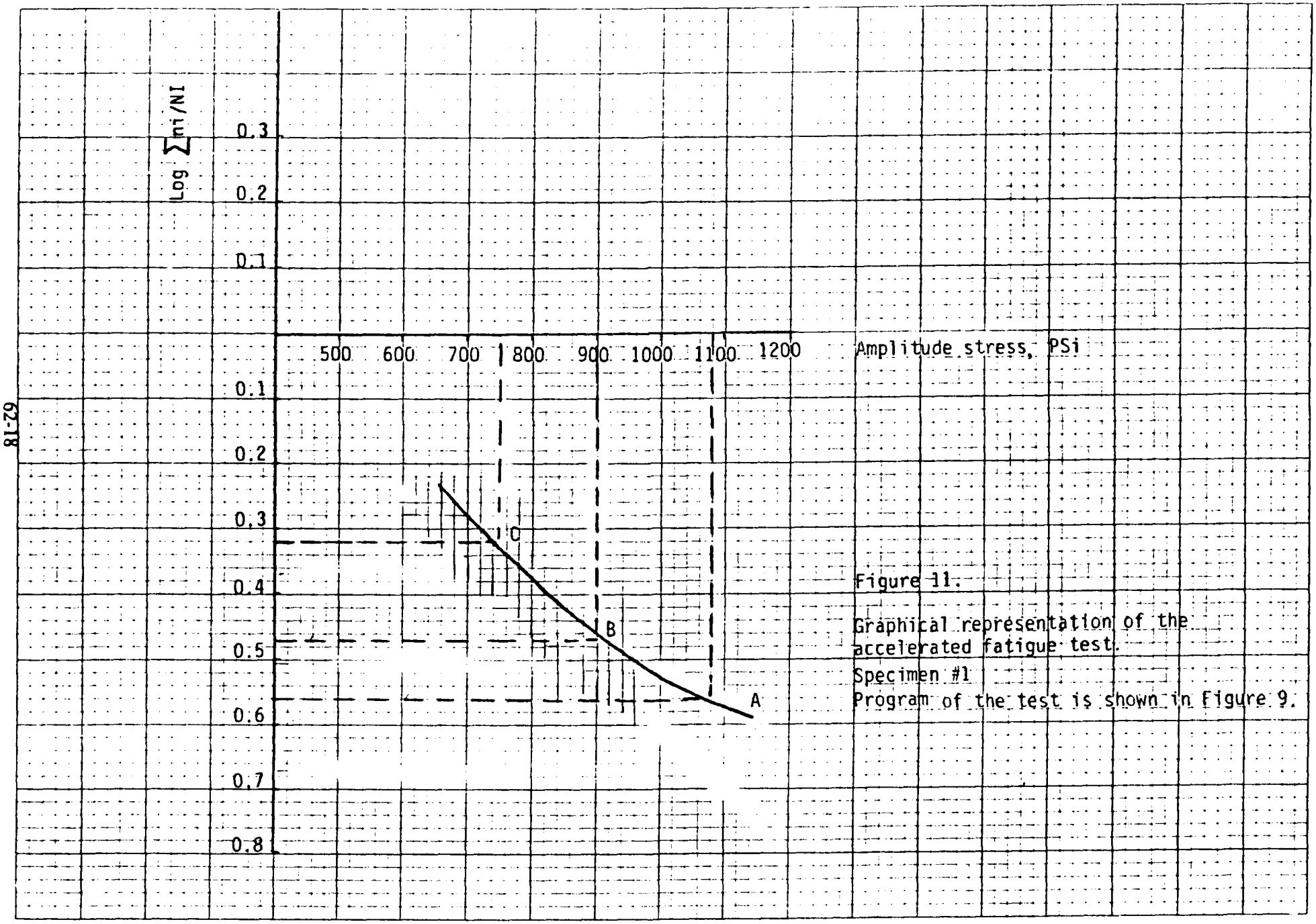


Figure 11.

Graphical representation of the accelerated fatigue test.

Specimen #1

Program of the test is shown in Figure 9.

ADDITIONAL FATIGUE TESTS

The investigation of the fatigue lives of the specimens with different thicknesses and the investigation of the testing frequency influence on a final fatigue life were conducted in addition to the major goal of the project (development of the accelerated fatigue test procedure).

Comparison of the fatigue lives for the 0.5" and o.25" thickness specimens tested under the same stress is given in Figure 12, The test result show that the life of the thinner material is less. It was observed that the cracks propagated with almost equal speed across the specimens and into their bodies. Therefore, taking into account that the width of the specimen from the hole edge to the side is 0.5", it is understandable that the test results are logical. The results could be important in the selection of the appropriate canopy material.

Table 4 reflects the results of fatigue tests conducted at different frequencies. The results show that the properties of the material tested are obviously time dependent. The number of cycles until complete separation strongly depends on frequency, bit specimen life in hours was almost the same for the specimens tested at the same load level and different frequencies. It provides an interesting observation about optimal usage and efficiency of the polycarbonate parts.

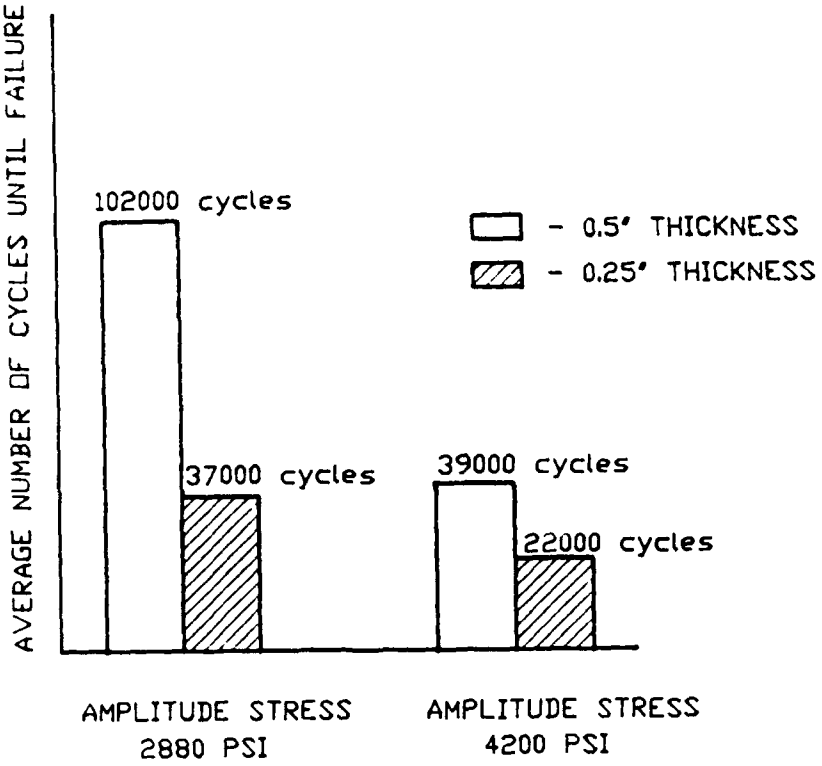


Figure 12. Comparison of fatigue lives of the specimens with different thickness. Variable bending. Specimen with stress concentration. Frequency 5 Hz. Five specimens tested in each group.

TABLE 4
Fatigue lives of specimens tested at different frequencies. Specimens with stress concentration. Amplitude load is 240 Lb.

Specimen number	Frequency, Hz	Number of cycles until failure	Time until failure, h	Average failure time, h
1	8	119300	4.14	4.69
2	8	151000	5.24	
3	5	92000	5.11	
4	5	99000	5.5	
5	5	115000	6.39	5.31
6	5	100000	5.57	
7	5	72000	4.01	
8	2	40900	5.68	
9	2	33700	4.68	4.70
10	2	61300	3.74	

CONCLUSIONS AND RECOMMENDATIONS

The project considered in this report is a continuation of the study started by the principal investigator at Flight Dynamics Laboratory, Wright-Patterson Air Force Base in 1989 and continued in 1990 and 1991 years. Therefore, the conclusions given below are overall conclusions based on the results given in this report and in [1] and [2].

1. The 0.5" thickness polycarbonate specimens have significant life until complete separation after appearance of the minute cracks.

2. The massive minute cracks usually appear only on the specimens tested at high load levels. Therefore, it is possible to make a certain judgment about load history by analyzing the specimens after failure.

3. The influence of a stress concentration on the fatigue life of the polycarbonate specimens tested is very strong. The specimens without stress concentration have much longer fatigue life. From this point of view, frameless canopies without holes for fasteners can be very promising. However, it is important to keep in mind that the scatter of the test results for the solid specimens is greater than that for the specimens with holes.

4. The results of the accelerated tests run have good repeatability and, a procedure used can be recommended to control the stability of manufacturing process. The procedure can be also useful, for example, for preliminary comparison of the material for

frameless canopy and conventional polycarbonate sheet which is used today.

5. The search of the optimal parameters of the accelerated test can be continued. The author believes that the testing time can be decreased.

6. The fatigue tests run proved that the fatigue strength of the polycarbonate sheets varies significantly from sheet to sheet and probably from manufacturer to manufacturer. Therefore, the quick control of fatigue characteristics can be recommended as a compulsory procedure.

7. The polycarbonate sheet thickness (at least in a range from 0.25" to 0.5") has very strong influence on the fatigue life of this material.

8. The properties of the specimens are strongly time dependent. The final specimen life in hours does not depend on the testing frequency.

9. The investigation of fatigue resistance of the canopy materials in different environmental conditions is strongly recommended.

REFERENCES

1. Yulian B. Kin, "Fatigue characteristics of F-16 composite transparency material determined by long-term and accelerated methods", Final Report, Contract No. FY 9620-88-0053/SB5881-0378, Universal Energy Systems, Inc., Dayton, Ohio, February 1990.

2. Yulian B. Kin, "Accelerated fatigue test procedure for the structural polycarbonate component of the F-16 canopy composite material", Final Report, Contract No. F49620-88-C-0053, Flight Dynamics Laboratory, Wright-Patterson Air Force Base, August 1990.

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
1991 RESEARCH INITIATION PROGRAM

Conducted by the
Universal Energy Systems, Inc.
4401 Dayton-Xenia Road
Dayton, Ohio 45432

FINAL REPORT

FATIGUE FRACTURE BEHAVIOR OF CORD-REINFORCED RUBBER COMPOSITES

Prepared by:	B. L. Lee and D. S. Liu
Academic Rank:	Associate Professor and Graduate Student
Department and	Engineering Science and Mechanics
University:	Pennsylvania State University, Univ. Park, PA
Research Location:	Pennsylvania State University, Univ. Park, PA

Date:	31 December, 1991
Contract No:	Purchase Order S-210-11M6-088

ABSTRACT

Interply shear fatigue behavior of nylon cord-reinforced rubber matrix composites has been investigated to assess the mechanisms of *local damage accumulation* and *delamination failure* of bias aircraft tire carcass. In the case of flat coupon specimens of finite width, interply shear strain was found to induce localized failure in the forms of *cord-matrix debonding* and *matrix cracking* which were eventually developed into the *delamination*. One unique feature of aircraft tire carcass composites was a relatively long period of time sustained after the onset of delamination. Fatigue lifetime profile of aircraft tire carcass composites was found to be dependent on the level of frequency (speed). A more dominant role of *creep* effect due to the viscoelastic properties of constituent materials is suspected for reduced fatigue lifetime at a lower frequency. The process of damage accumulation in aircraft tire carcass composites was accompanied by a continuous increase of *temperature* and *cyclic strain*. In determining the rate of cyclic strain increase for the composites, the effect of creep caused by viscoelastic properties of constituent materials appears to be as important as the effect of material flow and subsequent crack growth from the debonded cord ends. The fatigue life of aircraft tire carcass composites was inversely proportional to the *rate of cyclic strain increase*. However, the extent of *dynamic creep at gross failure*, which is defined as the increase of cyclic strain beyond initial elastic deformation, was roughly independent of stress amplitude. The use of internally-pressurized tube specimens of angle-plied composites was explored to simulate the delamination of aircraft tire carcass in shoulder region with no cut ends of cords. However, under the externally applied axial tension, the tubular specimens exhibited extensive interply shear deformation superseded by straining of cord reinforcement.

I. INTRODUCTION

Our research effort ultimately aims at the laboratory prediction of the *durability* of *aircraft tires* based on the study of deformation and fracture behavior of cord-rubber composite elements. The effort has been initiated for the Landing Gear Systems Group (LGSG) of the Flight Dynamics Directorate at Wright Research Laboratory to develop the capabilities of predicting long-term performance and life expectancy of aircraft tires in a cost-effective manner. Compared with other types of pneumatic tires, aircraft tires are subjected to unusual combinations of speed and load (*1,2*). For example, the baseline tire used by LGSG in their current durability study is a 49X17/26PR bias construction and is rated at a speed of 358 km/hr (224 mph) and a load of 173 500 N (39 000 lbf). Extreme combinations of speed and load result in high cyclic frequencies, large deflections and significan heat generation due to hysteretic loss in aircraft tires.

As confirmed by field experience (*3*), these conditions cause damage in critical sub-regions of tires such as *shoulder*, *bead*, *lower sidewall* or *tread* which eventually develops into catastrophic failures of whole tires. The failure of the carcass ply in the shoulder area, often called a ply separation, occurs in the form of *delamination* which involves crack propagation mainly in the rubber matrix and cord-matrix interface. So-called bead area cracking and lower sidewall break also involve crack propagation in the rubber matrix of carcass ply, but there are strong indications of *fiber fracture* as well. The processes of damage accumulation and structural failure of tire carcass in the shoulder, bead and lower sidewall areas are attributed to a combination of mechanical overloading and heat generation along with the resultant deterioration of constituent materials. On the other hand, tread groove cracking and subsequent tread-carcass separation of aircraft tires are attributed to the presence of strong centrifugal force resulting from unusually high speed particularly when so-called *standing wave* is present (*1-5*).

As discussed so far, basic understanding of tire failure mechanisms has been established at least in qualitative sense. In the case of *tread* failure, its major cause, the occurrence of standing waves can be avoided by a proper design of tires (*4*). However, as far as the *carcass* ply is concerned, the tasks of identifying critical operating conditions (speed, load, underinflation and tire deflection) responsible for the failure and predicting the useful life

expectancy of an aircraft tire are still difficult at best. Accelerated testing based on *dynamometers* provides a valuable means of evaluating the structural durability and life expectancy of aircraft tires. But it is too costly to rely solely on the dynamometer test results in specifying the number of takeoff and landing cycles a tire may endure before replacement (*1*). Besides the results of these dynamometer tests reflect merely the sensitivity of each specific tire design and construction to a specific combination of test conditions, unless underlying mechanisms of material property degradation and damage accumulation are identified.

Past approaches to characterize the property degradation profile of the aircraft tires have utilized coupon specimens cut from the carcass region (*3,6,7*). This type of durability analysis provides useful information on failures caused by the deterioration of material properties such as ply-to-ply adhesion strength. Some estimates assume that tire failure occurs at a 50 percent reduction of carcass ply adhesion strength. However, these works did not show how the deterioration of material properties leads to damage accumulation and eventual structural failure of tire carcass. Our research program plans to clarify these issues by investigating the *mechanisms of local damage accumulation, material property degradation and structural failure* for the cord-rubber composites which represent the aircraft tire carcass. In the study, the mechanisms of damage accumulation, material property degradation and structural failure of cord-rubber composites will be examined under the laboratory loading conditions which in turn simulate individual elements of complex loading for the actual tire carcass.

The above-described research program has begun in the 1990 Summer Faculty Research Project (SFRP) with an emphasis on failure modes in the *shoulder* area (*8*). In the 1990 SFRP, an experimental method was developed to simulate *carcass delamination* process in the flat coupon specimens of angle-plied cord-rubber composites. For a better determination of the failure modes, SFRP utilized mainly the *model* composite coupons reinforced by steel wire cables of large diameter. Under uniaxial cyclic loading which represents circumferential tension in the footprint region of bias tires, angle-plied cord-rubber composite specimens were found to exhibit extensive interply shear deformation which eventually leads to delamination type failures. Although the use of the coupon specimens with free edges allows a direct observation of the mechanisms of damage accumulation and delamination, the cut ends of cords exposed at the specimen edge also act as built-in defects in the process of damage initiation (*9,10*). This fact is expected to

complicate our future interpretation of composite test results in correlation with tire durability data.

As a follow-up study of SFRP, a current Research Initiation Program has expanded the scope of work to include detailed and more realistic assessment of interply shear fatigue behavior of the nylon cord-reinforced composite coupons which represent the actual *aircraft tire carcass*. The study has allowed initial screening of the operating parameters (e.g. frequency, stress amplitude, minimum cyclic stress, strain range, dynamic creep rate, temperature, heat-up rate, etc) which can influence the mechanisms of fatigue damage accumulation and delamination failure of the composites. The mechanism of material property degradation has not been fully investigated because of time limitations. Finally, in addition to the flat coupons, the use of internally-pressurized tube specimens of angle-plied composites was explored to simulate the condition of aircraft tire carcass which has no cut ends of cords in shoulder region but delaminates under biaxial tension.

II. OBJECTIVES OF THE RESEARCH EFFORT

(a) To assess interply shear deformation and fracture behavior of aircraft tire carcass composites under fatigue loading, (b) to identify the operating parameters which can influence the mechanisms of fatigue damage accumulation, material property degradation and delamination failure of angle-plied cord-rubber composites, and (c) to develop a new biaxial fatigue test method involving internally-pressurized tube specimens without the cut ends of cords.

III. EXPERIMENTS

As in the case of the 1990 SFRP, the current study used mainly *flat coupon* specimens press-molded from calendered ply stocks (Table 1). The specimen consists of a cord-rubber composite representing typical *bias aircraft tire carcass* using a +/- 38 deg reinforcement angle, 1260/2 nylon cord and a proprietary rubber compound based matrix. The aircraft tire carcass composite was found to have a lower modulus and a lower strength than the *model* composite which will be often quoted in this report as a reference system

(*8*). The model composite uses a +/- 19 deg cord angle and steel wire cables of 1.62mm diameter circular cross-section, because the large cord diameter and the cord angle chosen maximize the interply shear strain, a major contributing factor in composite failure. The composite coupon specimens had free edges with the cut ends of reinforcing cords exposed. To avoid tension-bending coupling, the coupon specimens were constructed with a symmetric ply lay-up. The end tabs were added to the specimens to prevent failure near the gripping region.

In simulation of the circumferential tension imposed to an aircraft tire in the footprint region, composite coupon specimens were subjected to both static and cyclic uniaxial tensile loading. Cyclic testing was performed under a broad range of stress amplitude and three different levels of *frequency* (1, 5 and 10 Hz) to generate a series of S-N (stress range vs number of cycles to failure) curves. Minimum cyclic stress was $k c \rho t$ constant at 0.2 ksi. Throughout the testing, heat generation due to hysteretic loss was closely monitored relying on the surface temperature of the specimens. In both static and cyclic testing, local strain was estimated by measuring the displacement of line markings drawn on the specimen edge.

Although a systematic variation was not attempted, the following four different *length-to-width ratios* were employed in testing of flat coupons of aircraft tire carcass composites: *5.3* (0.75 inch wide / 4 inch long or 1 inch wide / 5.25 inch long); *4* (1 inch wide / 4 inch long); *3.3* (0.75 inch wide / 2.5 inch long); *0.5* (2.5 inch wide / 1.25 inch long). Among them, the specimens of a length-to-width ratio of 0.5 were tested to observe failure mode of the composites when no single cord reinforcement remains ungripped.

In addition to the flat coupons, *tube* specimens with no cut ends of cords were prepared from an identical cord-rubber composite ply stock representing *bias aircraft tire carcass* (+/- 38 deg cord angle, 1260/2 nylon cord and a proprietary rubber compound based matrix). In this case, a symmetric ply lay-up was not used. Some tube specimens were prepared with a pre-delaminated area by inserting Teflon sheet between the plies. The composite specimens with or without pre-delamination were internally pressurized and tested under static tension along the longitudinal axis.

IV. RESULTS AND DISCUSSIONS

IV. 1. Failure Modes of Aircraft Tire Carcass Composites

As shown in our previous study (*8*) as well as numerous studies in the past (*9-13*), the angle-plied cord-rubber composites exhibit a large *interply shear strain* under static or cyclic tension. Interply shear strain develops in angle-ply laminates, because the constituent plies exhibit in-plane shear deformation of opposite direction but the action is prevented by mutual constraint due to interply bonding. Compared with the case of fiber-reinforced plastic composites (*14*), cord-rubber composites exhibit unusually high level of interply shear strain which results from the load-induced change of reinforcement angle allowed by extreme compliance of rubber matrix. The previously reported data indicate that, at an axial strain of say 10 percent under static tension, an interply shear strain of 70 percent develops in the nylon cord-reinforced aircraft tire carcass composites and 120 percent in the model composites (Figures 5 and 6 of Ref. *8*).

Such a large interply shear strain was found to readily induce localized failure in the form of *cord-matrix debonding* under cyclic as well as static tension. Debonding was started around the cut ends of reinforcing cords at the edge of the finite width coupons, which is justified since the maximum interply shear strain occurs at the edge of the specimen. A critical level of interply shear strain for cord-matrix debonding corresponded to the axial stress of approximately 20 to 30% of ultimate strength of the composites. Under static tension, cord-matrix debonding developed into *matrix cracking* as the strain increases. At higher strain, the axial stress-strain curves exhibited strain hardening type response for both aircraft tire carcass composites (Figure 1) and model composites (Figure 8 of Ref. *8*). Debonding and matrix cracking were widened with increasing strain and eventually developed into the *delamination* leading to gross fracture of the composites. The same sequence of failure modes was observed under cyclic loading as long as minimum stress remains in tensile regime.

Our unpublished results indicate that the described sequence of failure modes is reversed in the model composites when minimum cyclic stress is near zero. The delamination process appears to precede cord-matrix debonding in this situation. One interesting observation on the aircraft tire carcass composites was that the development into the delamination occurs when a *length-to-width (L/W) ratio* of the specimen is greater than 1.28 (Cot 38 deg) and therefore some of the reinforcing cords remain ungripped. In the case of aircraft tire

carcass composite specimen of a L/W ratio of 0.5, the absence of any ungripped reinforcing cords led to a final failure mode of *fiber fracture* instead of the delamination under static tension.

IV. 2. Fatigue Behavior of Aircraft Tire Carcass Composites and Frequency Effects

Since the minimum cyclic stress is kept at 0.2 ksi safely in tensile regime, all tested specimens (L/W = 3.3 to 5.3) of nylon cord-reinforced aircraft tire carcass composites exhibited a normal sequence of failure modes i.e. cord-matrix debonding developing into the delamination as in the model composites. One unique feature of aircraft tire carcass composites was a relatively long period of time sustained after the onset of partial delamination. This tendency was more pronounced at lower stress amplitudes. As a result, the S-N curve for aircraft tire carcass composites is linear on a log-log scale allowing the determination of a power law factor (Figures 2 and 3), while the S-N curve for the model composites is linear on a sem-log scale (Figure 9 of Ref. *8*). As in the case of the model composites, the critical load for the onset of cord-matrix debonding seems to constitute a threshold level for semi-infinite fatigue life, i.e. *endurance limit*, of the aircraft tire carcass composites. However, the presence of the endurance limit is less clear-cut, because prolonged fatigue life after the onset of delamination requires testing beyond 10^7 cycles.

The effect of frequency on the fatigue behavior of aircraft tire carcass composites has been studied using the specimens of three (eventually four) different L/W ratios as shown below.

		L/W Ratio			
		0.5	3.3	4.0	5.3
		1.25" L 2.5" W	2.5" L 0.75" W	4" L 1" W	4 or 5.25" L 0.75 or 1" W
Freq:	1 (Hz)	*Planned*	Complete*	Partly done	In progress
	5	-	Partly done*	-	Complete*
	10	*Planned*	In Progress	-	Partly done*
	20	-	*Planned*	-	*Planned*

Although testing has not been completed, fatigue lifetime profile of aircraft tire carcass composites seems to be dependent on the level of frequency (speed). Compared with the

case of 5 Hz, consistently shorter fatigue lifetime is observed under the frequency of 1 Hz at a given stress range or maximum stress (Figures 3 and 4). To a much lesser extent, shorter fatigue lifetime is also observed under the frequency of 10 Hz compared with the case of 5 Hz (Figures 5 and 6). If our future test results confirm a further reduction of fatigue lifetime at a higher frequency of 20 Hz, the phenomenon could be attributed to a greater rate of *heat* generation due to hysteretic loss. In contrast, a more dominant role of *creep* effect due to the viscoelastic properties of constituent materials, which will be discussed in the next section, is suspected for distinctly shorter fatigue lifetime at a lower frequency of 1 Hz. Static creep loading experiments are planned to assess the contribution of viscoelastic properties of constituent materials to the damage accumulation process of composites.

IV. 3. Fatigue Damage Accumulation Process of Aircraft Tire Carcass Composites

As observed in the model composites (Figure 10 of Ref. *8*), the process of damage accumulation in aircraft tire carcass composites is accompanied by a continuous increase of *temperature* and *cyclic strain*. The change of cyclic strain (either maximum strain or strain range) undergoes three stages after an initial stepwise increase. At first, the strain increases rapidly but at a progressively lower rate. When plotted against the logarithm of the time, this first region often forms a linear part of a sigmoidal shape curve (Figures 7, 8 and 9). In the second region, the cyclic strain increases at a constant rate. The region forms an increasingly steeper (upward curvature) portion of sigmoidal shape curve when plotted against the logarithm of the time. Throughout the first and second regions which consume a major part of fatigue life, the damage accumulation occurs in the forms of cord-matrix debonding and matrix cracking. Towards the end of the second region, partial delamination appears at the specimen edge. In the final third region, the cyclic strain increases rapidly at a progressively higher rate eventually leading to a catastrophic failure of gross delamination.

The first region of the cyclic strain vs time curve was found to be considerably longer in the aircraft tire carcass composites compared with the case of model composites. In the model composites, the first region takes less than 20% of the fatigue lifetime. On the whole, the curves of cyclic strain vs time for both composites strikingly resemble a static creep curve of four-element spring-dashpot model for typical polymeric materials (Figures 10 and 11) (*15*). The model postulates that the steepest part of the sigmoidal shape curve (i.e. the first

region) occurs within a period of time around the retardation time. At periods much longer than the retardation time, the only response to stress is believed to be due to the viscous flow. Therefore the prolonged presence of the first region suggests that, in determining the rate of cyclic strain increase for nylon cord-reinforced aircraft tire carcass composites, the effect of creep caused by viscoelastic properties of constituent materials is as important as the effect of material flow and subsequent crack growth from the debonded cord ends.

As shown in Figures 7-9, the fatigue life of aircraft tire carcass composites was found to be inversely proportional to the *rate of cyclic strain increase* (referred as *dynamic creep rate* in our previous study). The fatigue life of composites was also found to be inversely proportional to the *cyclic strain at gross failure* (Figures 12, 13 and 14), although some difficulty is encountered in establishing a clear trend with the maximum cyclic strain rather than the strain range. One interesting fact was that the extent of *dynamic creep at gross failure*, which is defined as the increase of cyclic strain beyond initial elastic deformation, is roughly independent of stress range. Under the frequency of 1 Hz, most specimens were found to fail when the maximum cyclic strain reaches approximately 30% above the level imposed by initial stepwise increase in strain (Figure 15). Although it may be premature to propose a formal criterion without additional data at different levels of frequency, the observation strongly suggests the existence of a critical level of dynamic creep for gross failures of the angle-plied cord-rubber composites and possibly for the tire carcass.

Although more data are needed before drawing a final conclusion, it appears that the correlation between the fatigue life and the rate of cyclic strain increase is independent of the frequency and therefore provides a universal means of life prediction (Figures 12, 13 and 14). However, the correlation between the fatigue life and the rate of temperature increase seems to be dependent upon the frequency used (Figures 16, 17 and 18). These facts suggest that, unless very high frequency (above 10 Hz) loading is applied, the origins of dynamic creep and damage accumulation of composites may be predominantly *mechanical fatigue* rather than thermal fatigue. In our future work, a more accurate determination of temperature distribution throughout the specimen will be attempted by embedded thermocouple technique. Aside from these questions on the roles in fracture mechanisms, the results demonstrate the measurement of local strain change and heat generation as a viable experimental technique for real-time monitoring of the damage accumulation process.

IV. 4. Development of Test Methodologies for Composite Tube

In the current study of aircraft tire carcass composites, the use of the coupon specimens with free edges allows a direct observation of the processes of damage accumulation and delamination failure. At the same time, monitoring of the line markings drawn on the specimen edge provides a straightforward means of determining interply shear strain under circumferential tension. However, the cut ends of cords exposed at the specimen edge act as built-in defects in the process of damage initiation. Besides the variation of meridional tension due to the inflation pressure cannot be simulated in the case of flat coupon specimens under uniaxial loading. Considering the above-described limitations, a new specimen configuration of internally-pressurized tube of cord-rubber composite was designed with an internal diameter of 0.5 inch and a gage length of 4 inches.

Under the externally applied axial tension, the air-inflated tube specimens exhibited extensive interply shear deformation and strain hardening behavior as in the case of flat coupons (Figures 19 and 20). One interesting observation was that, even with the presence of pre-delaminated area, straining of reinforcing cords eventually becomes a dominant mode of deformation leading to the slippage of the specimen from the grips. Since any ungripped reinforcing cords do not exist in the tube specimen subjected to axial tension (regardless of gage length variation), the prevention of grip slippage if possible is expected to lead to the fracture of reinforcing cords rather than delamination. As discussed previously, in the case of aircraft tire carcass composite coupon of a L/W ratio of 0.5 which is less than 1.28 (Cot 38 deg), the absence of any ungripped reinforcing cords led to a final failure mode of *fiber fracture* instead of the delamination under static tension.

The initial test results involving the inflated tube specimens of aircraft tire carcass composites clearly indicate that the occurrence of delamination cannot be simulated under static tension applied in the axial direction. Therefore our future study will be concentrated in the use of *cyclic tension* and/or *out-of-plane bending* mode to induce localized delamination and to avoid straining of the reinforcing cords as a dominant mode of deformation. The use of out-of-plane bending mode of deformation for the inflated tube specimen will become essentially an upgrading of so-called Mallory tube testing (*16*) which has been used mainly for the purpose of product quality control. The assessment of deformation and fracture behavior of carcass composites in biaxial stress will certainly complement our continuing research work of observing damage accumulation process of composites under simpler uniaxial tension.

V. CONCLUDING REMARKS

The described research effort ultimately aims at the laboratory prediction of the *durability* of *aircraft tire* based on the study of deformation and fracture behavior of cord-rubber composite elements. As a major step toward the goal, our research program plans to investigate the *mechanisms of local damage accumulation, material property degradation and structural failure* of the angle-plied cord-rubber composites which represent the aircraft tire carcass. These mechanisms will be studied under the laboratory loading conditions which in turn simulate individual elements of complex loading for the tire carcass. The program has begun in the 1990 Summer Faculty Research Project (SFRP) with an emphasis on *carcass delamination* process in the *shoulder* area using the model composites. As a follow-up study of SFRP, the current Research Initiation Program has expanded the scope of work to include detailed assessment of interply shear fatigue behavior of nylon cord-reinforced composites.

In angle-plied cord-rubber composites, unusually high level of interply shear strain was found to readily induce localized failure in the form of *cord-matrix debonding* under cyclic as well as static tension. As the strain increases, cord-matrix debonding developed into *matrix cracking*. Debonding and matrix cracking were further widened and eventually developed into the *delamination* leading to gross fracture of the composites. One unique feature of aircraft tire carcass composites was a relatively long period of time sustained after the onset of delamination. As in the case of the model composites, the critical load for the onset of cord-matrix debonding seemed to constitute the *endurance limit* of the aircraft tire carcass composites. Fatigue lifetime profile of aircraft tire carcass composites was dependent on the level of frequency (speed). Compared with the case of 5 Hz, consistently shorter fatigue lifetime was observed under the frequency of 1 Hz at a given stress level. A more dominant role of *creep* effect due to the viscoelastic properties of constituent materials is suspected for reduced fatigue lifetime at a lower frequency.

The process of damage accumulation in aircraft tire carcass composites is accompanied by a continuous increase of *temperature* and *cyclic strain*. The change of cyclic strain undergoes three stages after an initial stepwise increase. At first, the strain increases rapidly but at a progressively lower rate. In the second region, the cyclic strain increases at a constant rate. Throughout the first and second regions which consume a major part of fatigue life, the damage accumulation occurs in the forms of cord-matrix debonding and matrix cracking. Towards the end of the second region, partial delamination appears. The first region of the

cyclic strain vs time curve was found to be considerably longer in the aircraft tire carcass composites. The fact suggests that, in determining the rate of cyclic strain increase for the composites, the effect of creep caused by viscoelastic properties of constituent materials is as important as the effect of material flow and subsequent crack growth from the debonded cord ends.

The fatigue life of aircraft tire carcass composites was found to be inversely proportional to the *rate of cyclic strain increase*. One interesting fact was that the extent of *dynamic creep at gross failure*, which is defined as the increase of cyclic strain beyond initial elastic deformation, is roughly independent of stress range. Although it may be premature to propose a formal criterion without additional data at different levels of frequency, the observation strongly suggests the existence of a critical level of dynamic creep for gross failures of angle-plied cord-rubber composites and possibly for the tire carcass. The correlation between the fatigue life and the rate of cyclic strain increase is independent of the frequency and therefore provides a universal means of life prediction. However, the correlation between the fatigue life and the rate of temperature increase seems to be more dependent upon the frequency used.

Finally, in addition to the flat coupons, the use of internally-pressurized *tube* specimens of angle-plied composites was explored to simulate the condition of aircraft tire carcass which has no cut ends of cords in shoulder region but delaminates under biaxial tension. Under the externally applied axial tension, the tubular specimens exhibited extensive interply shear deformation and strain hardening behavior as in the case of flat coupons. However, straining of reinforcing cords eventually became a dominant mode of deformation since any ungripped reinforcing cords do not exist in the tubular specimen subjected to axial tension. Testing of flat coupon of a very small length-to-width ratio demonstrated that the absence of any ungripped reinforcing cords leads to a final failure mode of *fiber fracture* instead of the delamination.

VI. RECOMMENDATIONS

The following recommendations can be made for our continuing study on the deformation and fracture behavior of nylon cord-reinforced aircraft tire carcass composites:

(1) Continue fatigue testing until statistical confidence in the fatigue lives is obtained. Define the S-N curves for the onset of cord-matrix debonding and delamination as well as gross fracture.

(2) Continue the study of damage accumulation mechanisms with a systematic variation of *frequency*, *temperature* and the *ratio of minimum-to-maximum stress*. Perform a microscopic examination of damage accumulation modes.

(3) Confirm the existence of a critical level of dynamic creep at gross failure for aircraft tire carcass composites. Check the degree of reversibility of local strain encountered during dynamic creep.

(4) Assess the contribution of viscoelastic properties of constituent materials to the dynamic creep process of composites by performing static loading experiments at various temperatures. Determine how the process of damage accumulation interacts with material property changes.

(5) Define the exact role of hysteretic heating in determining the fatigue lifetime of composites by performing *isothermal* cyclic testing. Cyclic testing in isothermal condition will resolve the issue of the possible interaction between hysteretic heating and the progressive increase of strain due to damage accumulation.

(6) Assess the failure modes of internally-pressurized composite tube under *cyclic tension* and/or *out-of-plane bending* mode to induce localized delamination and to avoid straining of the reinforcing cords as a dominant mode of deformation.

REFERENCES

(*1*) S. N. Bobo, "Fatigue Life of Aircraft Tires", *Tire Science and Technology*, Vol. 16, No. 4, p.208 (1988).

(*2*) J. H. Champion and P. M. Wagner, "A Critical Speed Study for Aircraft Bias Ply Tires", AFWAL-TR-88-3006 (1988).

(*3*) Personal communications with the researchers in tire industry.

(*4*) J. Padovan, "On Standing Waves in Tires", *Tire Science and Technology*, Vol. 5, No. 2, p.83 (1977).

(*5*) J. H. Champion, S. K. Clark and M. K. Hilb, "A Study of Vibrational Modes in Rolling Aircraft Tires", WRDC-TR-89-3092 (1989).

(*6*) S. K. Clark, "Loss of Adhesion of Cord-Rubber Composites in Aircraft Tires", *Tire Science and Technology*, Vol. 14, No. 1, p.33 (1986).

(*7*) North Carolina A&T State University, "Study of the Behavior of Aircraft Tire Coupons under Various Loading Conditions", Report on Contract No. F33615-87-C-3413, U.S. Department of Air Force, Air Force Systems Command, Wright-Patterson AFB, OH (1990).

(*8*) B. L. Lee, J. P. Medzorian, P. M. Fourspring, G. J. Migut, M. H. Champion, P. M. Wagner and P. C. Ulrich, "Study of Fracture Behavior of Cord-Rubber composites for Lab Prediction of Aircraft Tire Durability", *Soc. of Automotive Engineers Paper* #901907, Warrendale, PA (1990).

(*9*) R. F. Breidenbach and G. J. Lake, "Mechanics of Fracture in Two-Ply Laminates", *Rubber Chemistry and Technology*, Vol. 52, p.96 (1979).

(*10*) R. F. Breidenbach and G. J. Lake, "Application of Fracture Mechanics to Rubber Articles Including Tyres", *Philosophical Trans. Royal Soc. London*, Vol. A299, p.189 (1981).

(*11*) J. D. Walter, "Cord-Reinforced Rubber" in Mechanics of Pneumatic Tires edited by S. K. Clark, U.S. Department of Transportation, Washington D.C. (1982).

(*12*) J. L. Ford, H. P. Patel and J. L. Turner, "Interlaminar Shear Effects in Cord-Rubber Composites", *Fiber Science and Technology*, Vol. 17, p.255 (1982).

(*13*) R. J. Cembrola and T. J. Dudek, "Cord/Rubber Material Properties", *Rubber Chemistry and Technology*, Vol. 58, p.830 (1985).

(*14*) Interlaminar Response of Composite Materials edited by N. J. Pagano, Composite Materials Series Vol. 5, Elsevier (1989).

(*15*) L. E. Nielson, Mechanical Properties of Polymers, Chap.3, Reinhold (1962).

(*16*) T. Takeyama, J. Matsui and M. Hijiri, "Tire Cord and Cord to Rubber Bonding" in Mechanics of Pneumatic Tires edited by S. K. Clark, U.S. Department of Transportation, Washington D.C. (1982).

Acknowledgements

We wish to thank the Air Force Office of Scientific Research for sponsorship of this research. Universal Energy Systems, Inc. helped us in all administrative aspects of this program.

We sincerely appreciate continuing support and encouragement for this research work from Messrs. John P. Medzorian, Paul C. Ulrich and Aivars V. Petersons and Dr. Arnold H. Mayer at the Vehicle Subsystems Division of Flight Dynamics Directorate of Wright Laboratory.

We are also very grateful to Dr. Alfredo G. Causa, Dr. Yao M. Huang and Mr. Ronald Burgan at Goodyear Tire & Rubber Co. for providing composite specimens.

TABLE 1

Specifications of Aircraft Tire Carcass Composites
(Cord: 1260/2 Nylon cord)
(Matrix: Proprietary rubber compound)

Cord Angle	-38, +38, +38, -38 deg
Cord Modulus	2.07 GPa (300×10^3 psi)
Matrix Modulus	5.51 MPa (800 psi)
Cross-Sectional Area of Cord	0.342 mm^2 (5.3×10^{-4} $inch^2$)
Total Width of Specimen	25.4 mm (1 inch)
Total Thickness of Specimen	6.35 mm (0.25 inch)

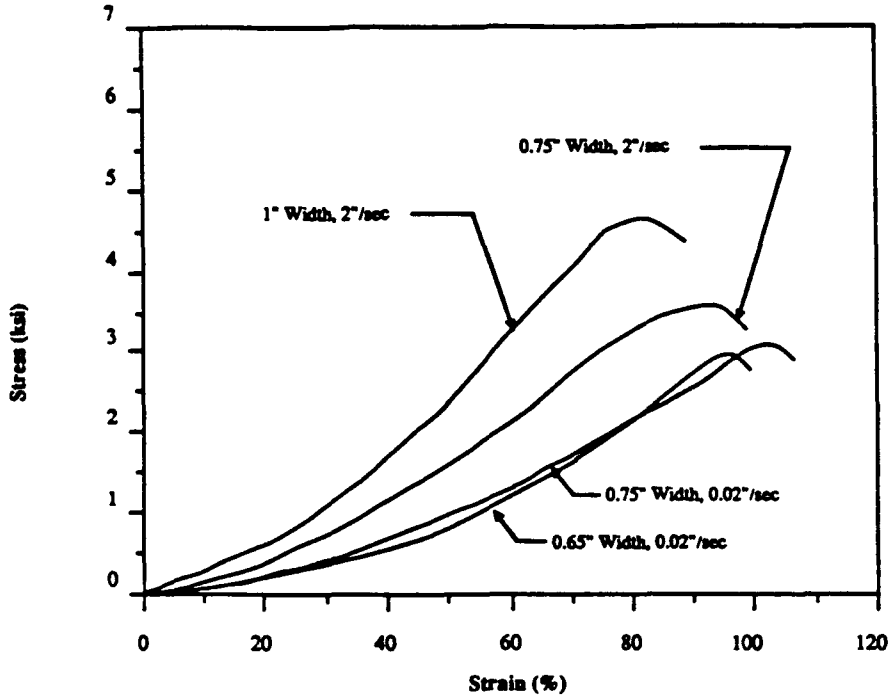


Figure 1

Stress-Strain Curves for Aircraft Tire Carcass Composites

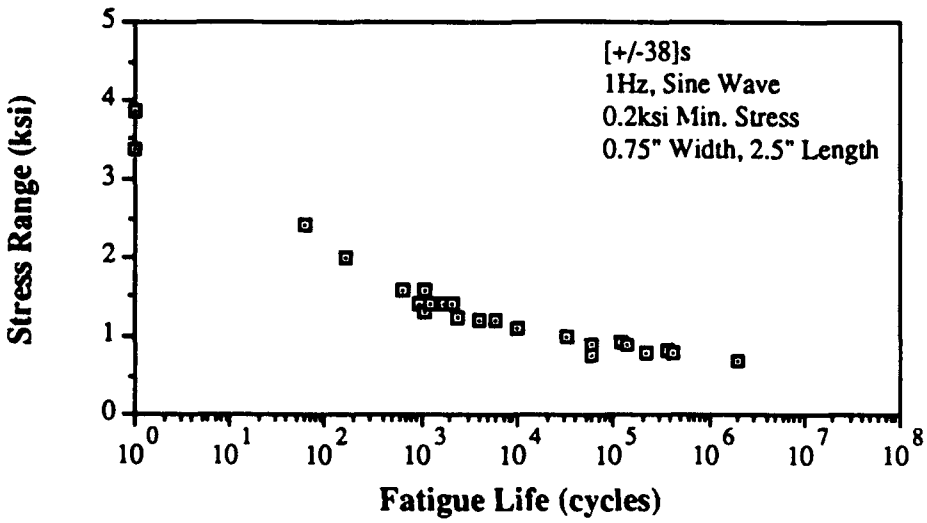


Figure 2

Stress Range vs Fatigue Lifetime (S-N) Curve for Aircraft Tire Carcass Composites
(1 Hz Frequency, L/W = 3.3, Semi-Log Scale)

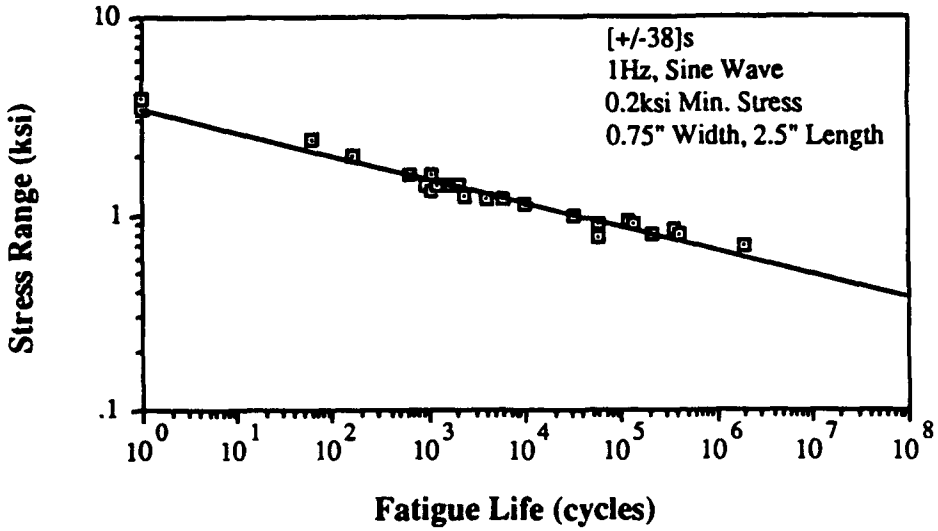


Figure 3

Stress Range vs Fatigue Lifetime (S-N) Curve for Aircraft Tire Carcass Composites
(*1 Hz* Frequency, L/W = 3.3, Log-Log Scale)

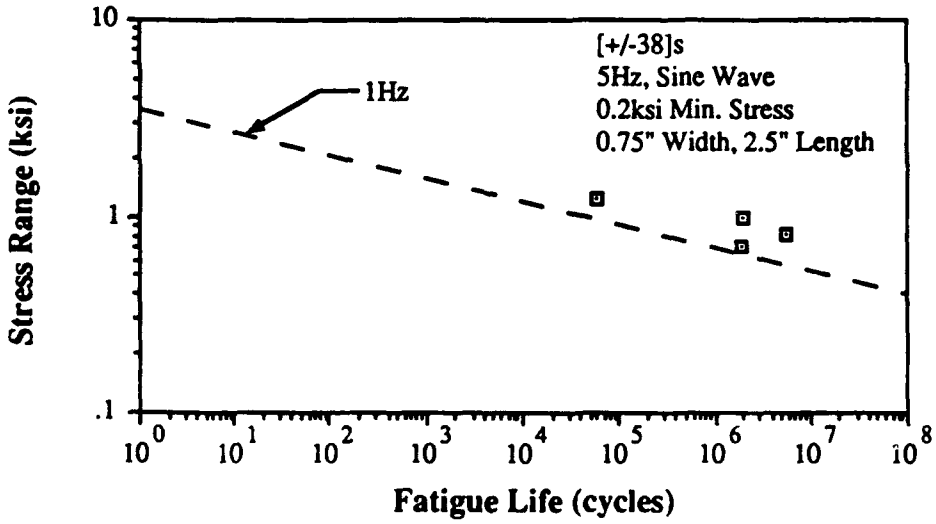


Figure 4

Stress Range vs Fatigue Lifetime (S-N) Curve for Aircraft Tire Carcass Composites
(*5 Hz* Frequency, L/W = 3.3)

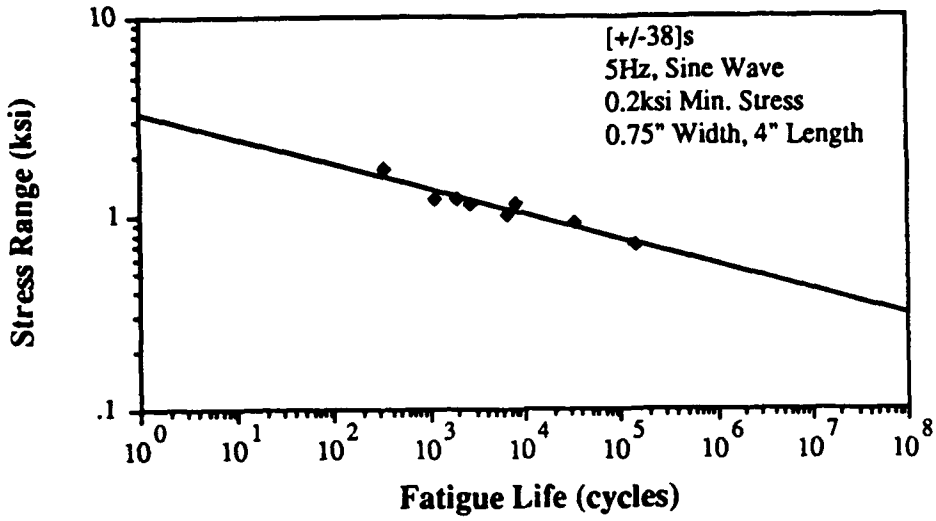


Figure 5

Stress Range vs Fatigue Lifetime (S-N) Curve for Aircraft Tire Carcass Composites
(*5 Hz* Frequency, L/W = 5.3)

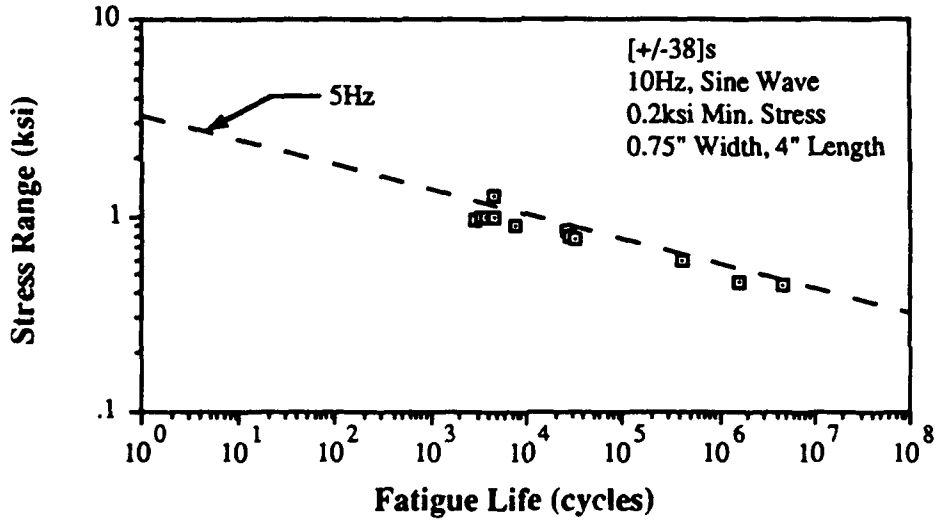


Figure 6

Stress Range vs Fatigue Lifetime (S-N) Curve for Aircraft Tire Carcass Composites
(*10 Hz* Frequency, L/W = 5.3)

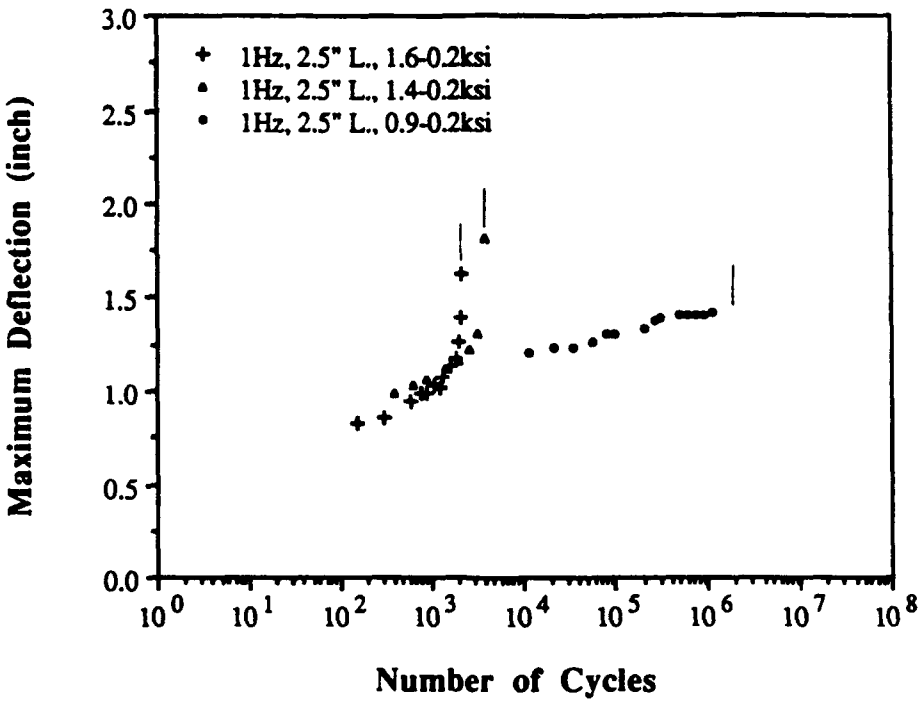


Figure 7

Maximum Cyclic Deflection vs Number of Cycles (*1 Hz* Frequency)

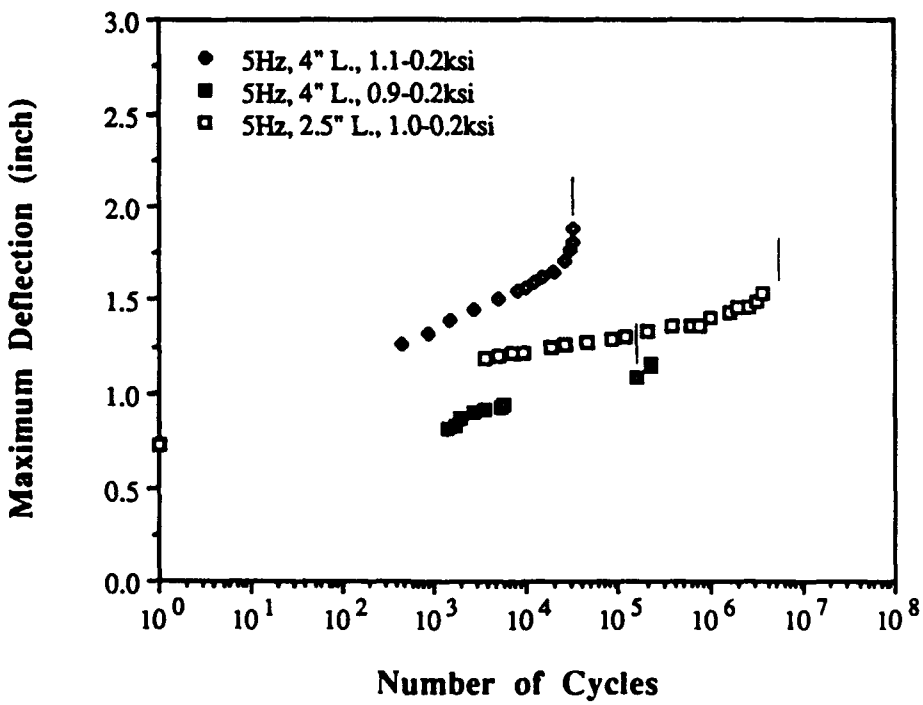


Figure 8

Maximum Cyclic Deflection vs Number of Cycles (*5 Hz* Frequency)

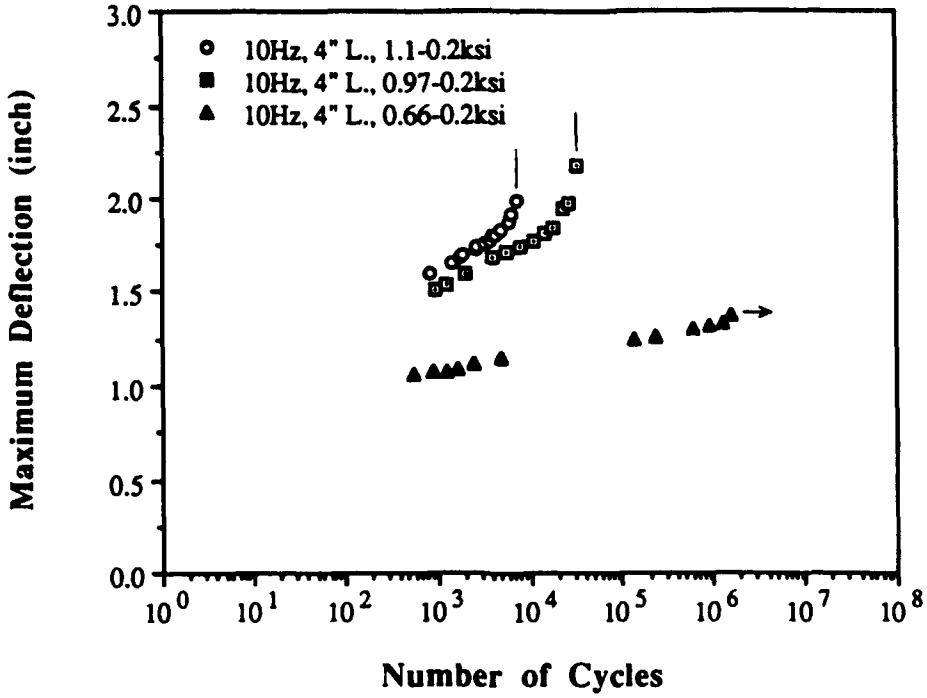


Figure 9

Maximum Cyclic Deflection vs Number of Cycles (*10 Hz* Frequency)

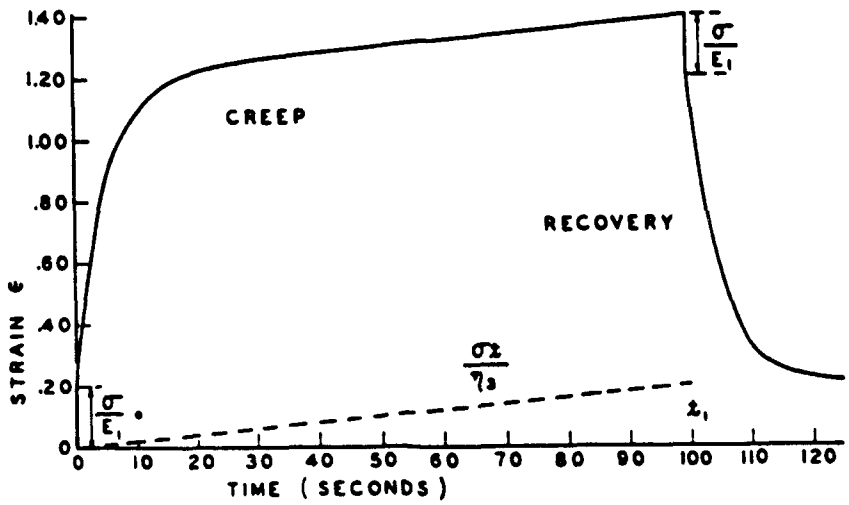


Figure 10

Creep of a Four-Element Model Plotted on a Linear Time Scale (from Ref. *15*)

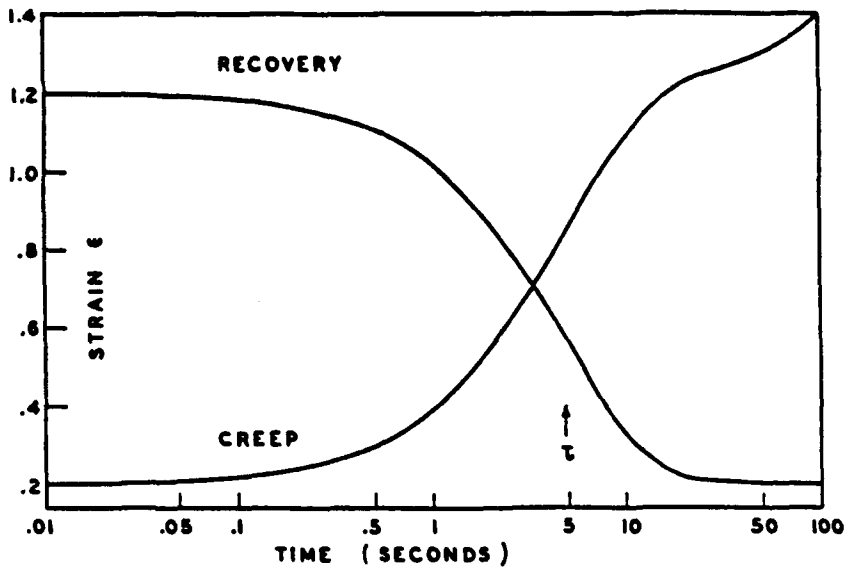


Figure 11

Creep of a Four-Element Model Plotted on a Log Time Scale (from Ref. *15*)

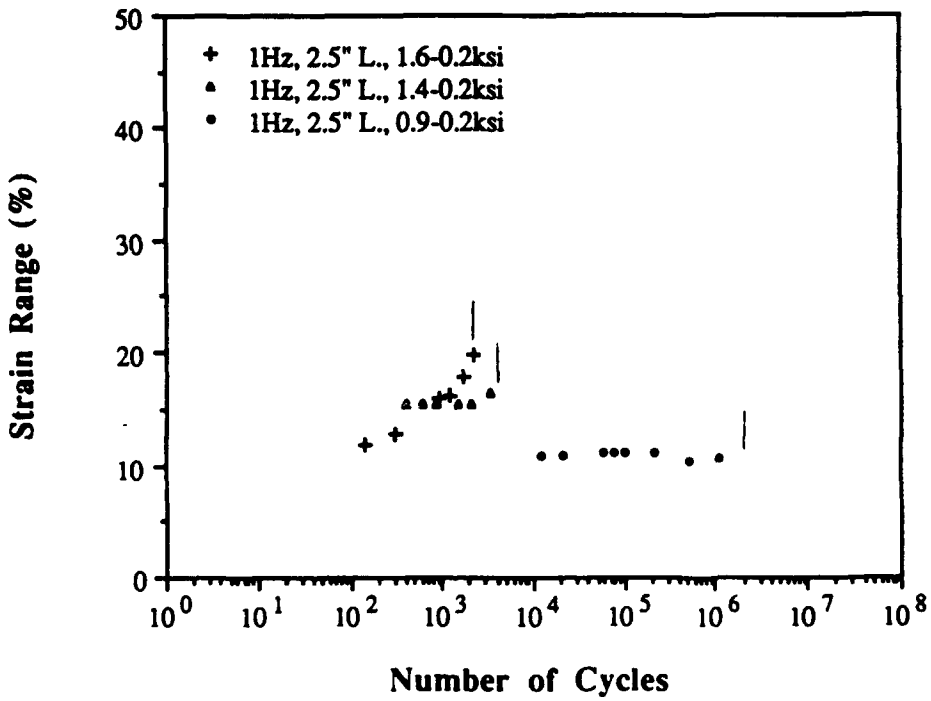


Figure 12

Cyclic Strain Range vs Number of Cycles (*1 Hz* Frequency)

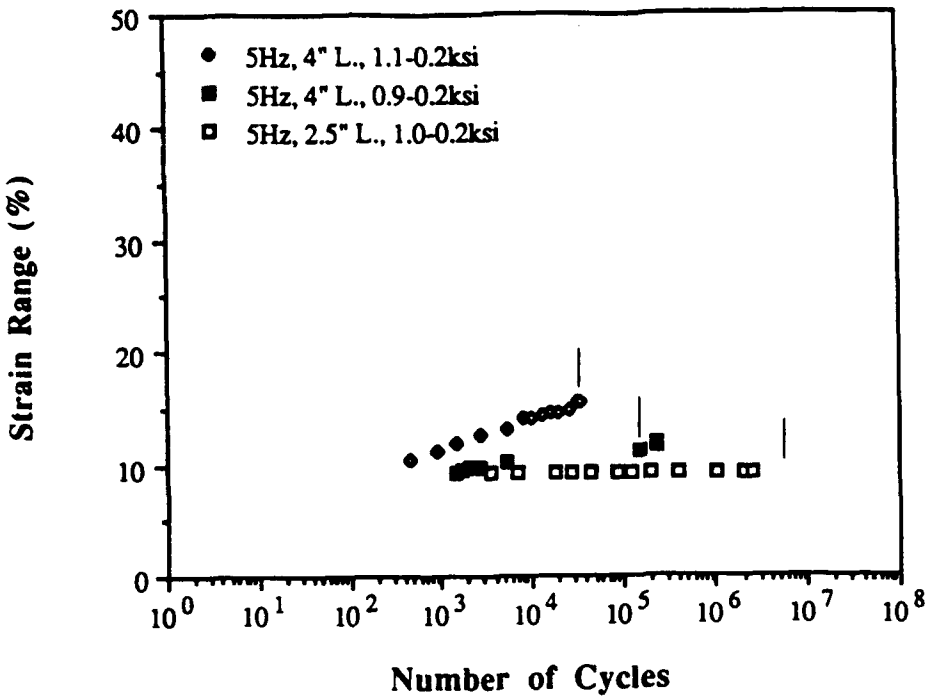


Figure 13

Cyclic Strain Range vs Number of Cycles (*5 Hz* Frequency)

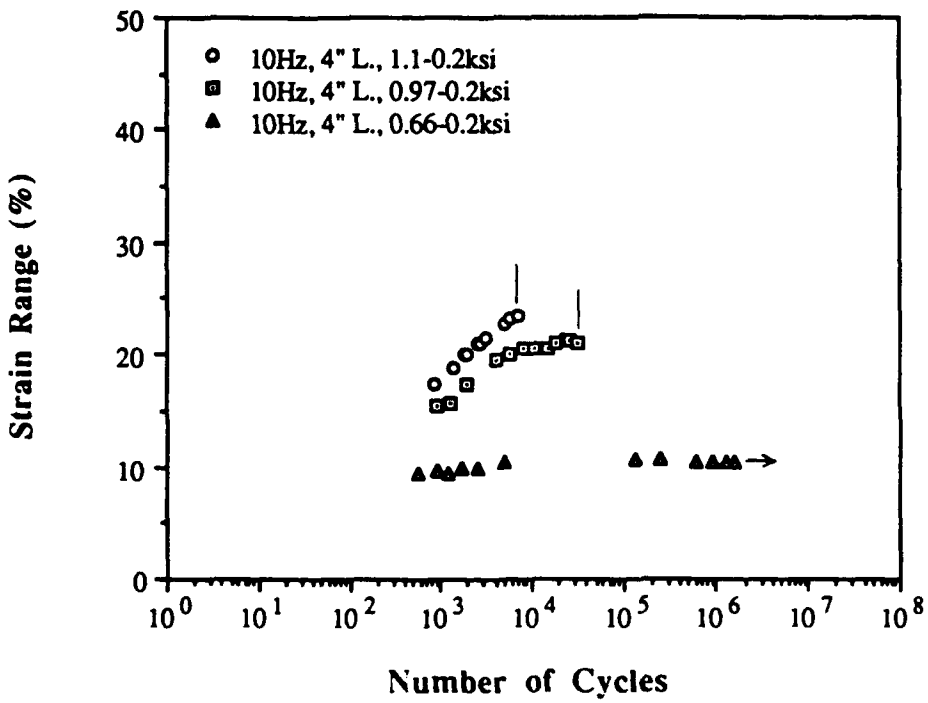


Figure 14

Cyclic Strain Range vs Number of Cycles (*10 Hz* Frequency)

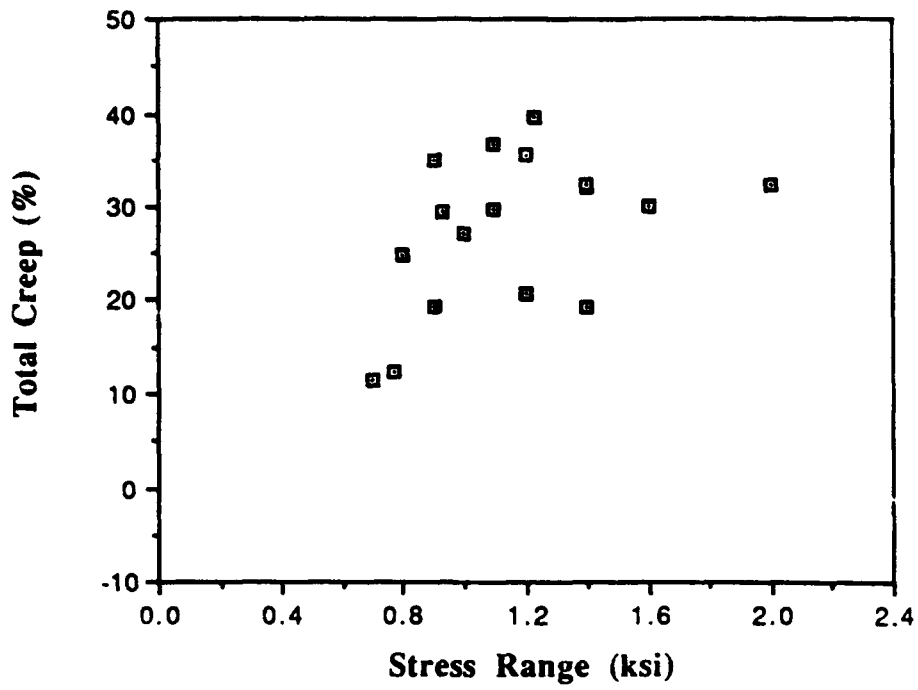


Figure 15

Dynamic Creep at Gross Failure vs Stress Range (*1 Hz* Frequency)

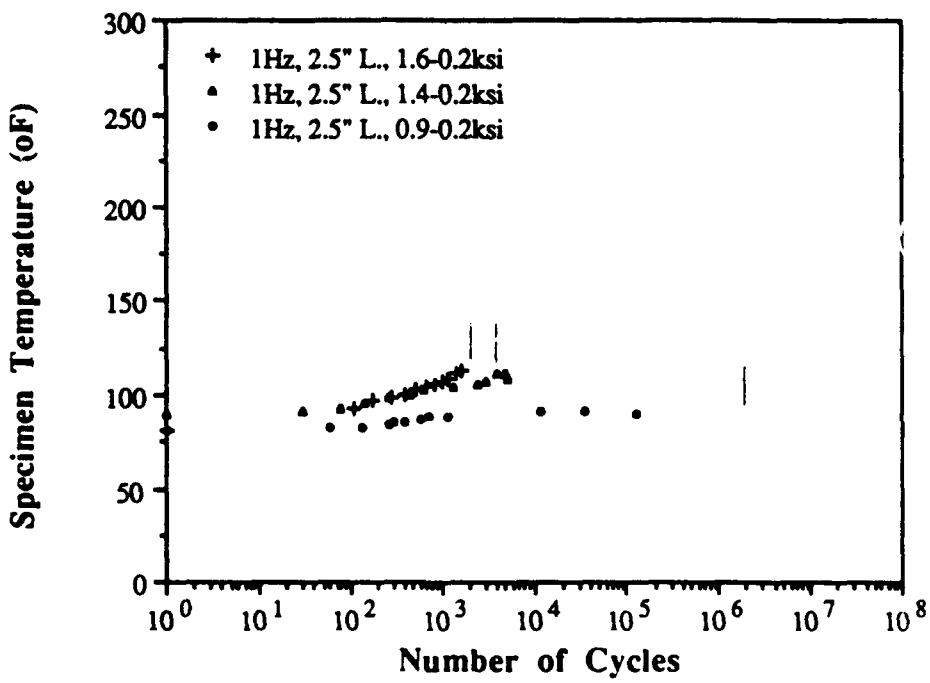


Figure 16

Specimen Temperature Rise vs Number of Cycles (*1 Hz* Frequency)

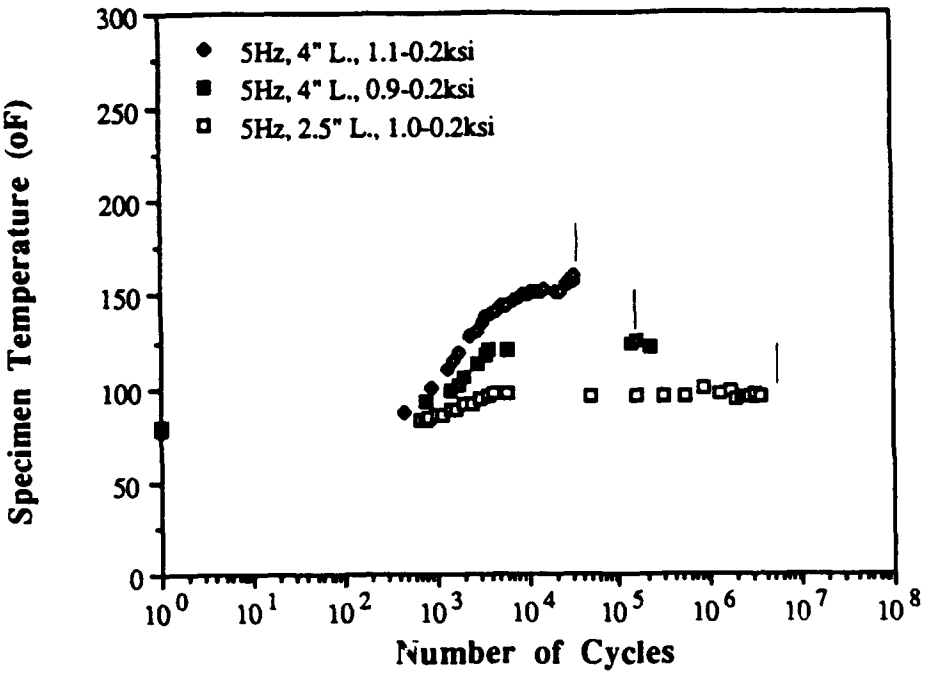


Figure 17

Specimen Temperature Rise vs Number of Cycles (*5 Hz* Frequency)

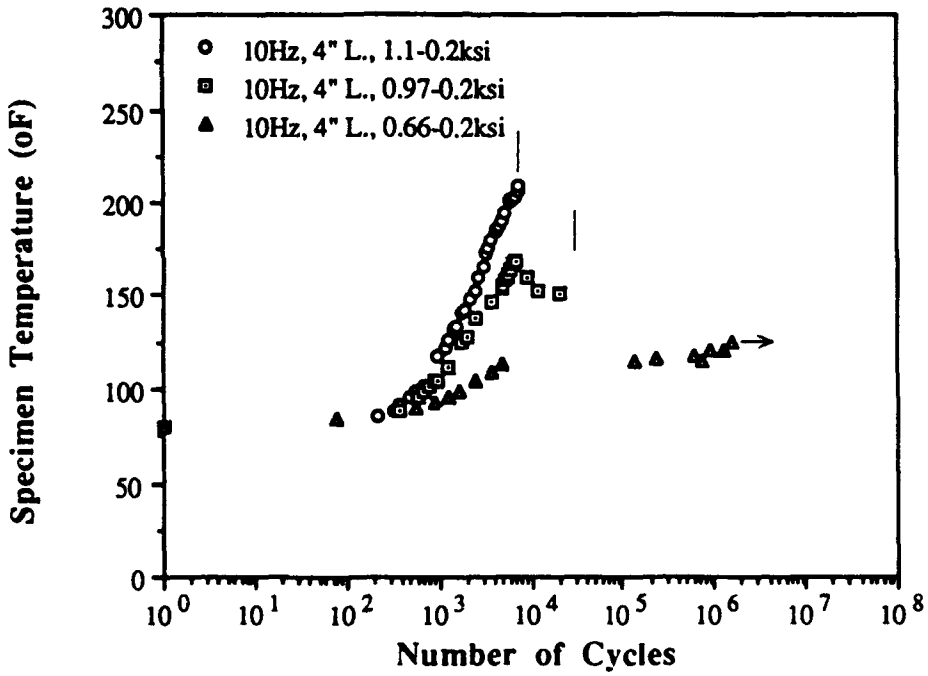


Figure 18

Specimen Temperature Rise vs Number of Cycles (*10 Hz* Frequency)

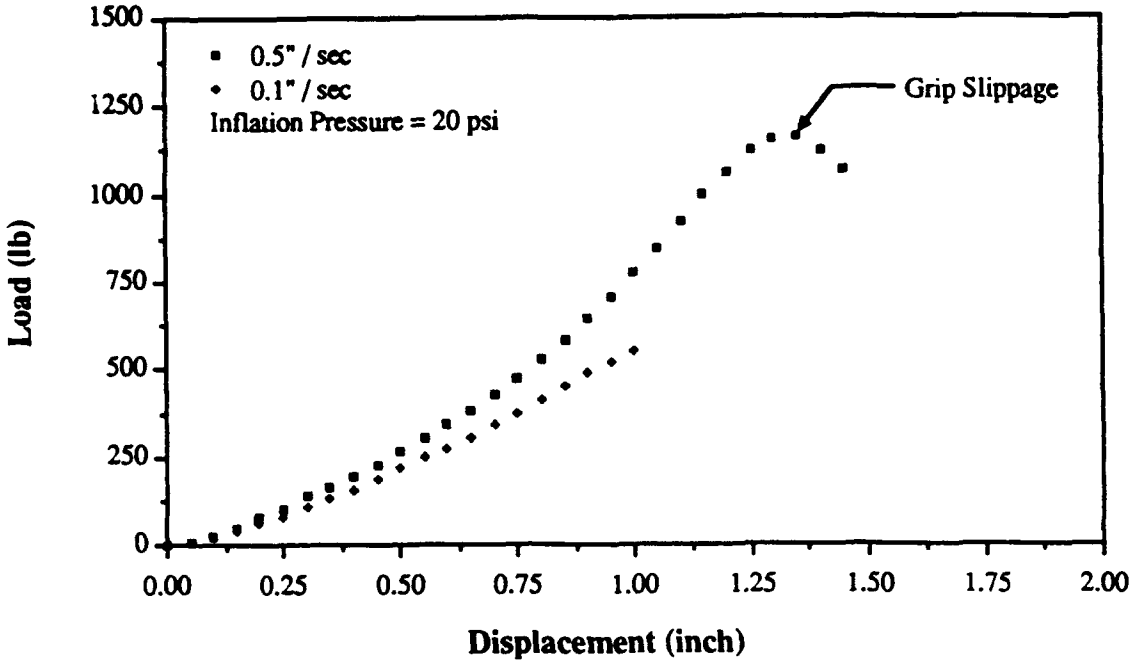


Figure 19

Load-Deflection Curve for Composite Tube Specimen

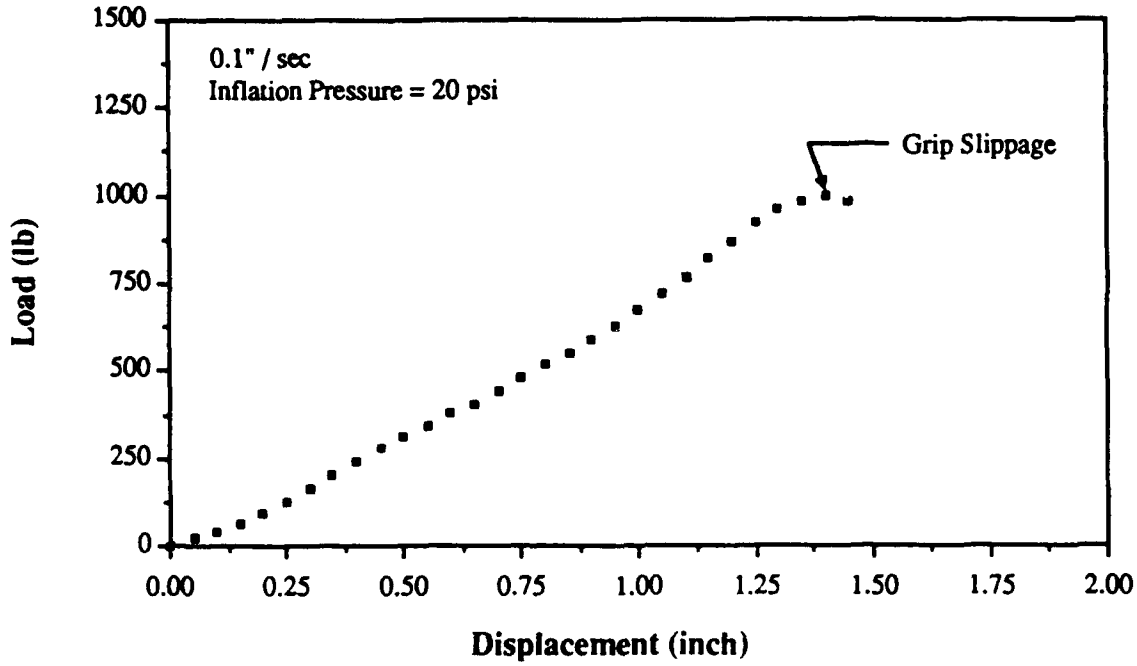


Figure 20

Load-Deflection Curve for Composite Tube Specimen with Pre-Delamination

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

Universal Energy Systems, Inc.

FINAL REPORT

THE EFFICACY OF CONSTRAINED-LAYER DAMPING TREATMENT TO SUPPRESS PARAMETRIC AND AUTOPARAMETRIC RESONANCES IN NONLINEAR AND INTERNALLY RESONANT NONLINEAR STRUCTURES

Prepared by:	Lawrence D. Zavodney
Academic Rank:	Assistant Professor
Department:	Engineering Mechanics
University:	The Ohio State University
Research Office:	Wright Laboratory Flight Dynamics Laboratory Structures Division Structural Dynamics Branch Acoustic and Sonic Fatigue Group

1.0 ABSTRACT

The influence of viscoelastic constrained-layer damping treatment on parametric resonances of single-degree-of-freedom (SDOF) systems and autoparametric resonances of multiple-degree-of-freedom (MDOF) nonlinear systems possessing autoparametric coupling was investigated. The results show that commercially available aluminium-backed treatment is effective in suppressing parametric resonances in SDOF systems by moving the regions of parametric resonance away from the frequency axis. In the MDOF systems the damping affects the highly nonlinear response characteristics of each mode and the nonlinear coupling between modes; it can suppress the modulation between modes. In general, the effect of increased application of damping treatment is to contract the regions of nonlinear modal interaction and, in some cases, actually suppress the nonlinear modal coupling entirely with a sufficient amount of damping treatment. Experimental results include slow swept-sine excitations at constant amplitude and slow swept-amplitude excitations at constant frequency. Particular attention was paid to the nonlinear resonances and the modal interaction regions bounded by the Hopf bifurcation.

2.0 INTRODUCTION

Parametric resonances are not uncommon in structural vibrations. Zavodney[1] and Zavodney, et al[2-4] have provided a summary of parametric resonance in SDOF nonlinear structural systems. Damping is a common method of controlling or reducing vibration for externally excited resonances. However, damping plays quite a different role in parametrically excited systems. When systems exhibit nonlinear behavior, the analysis is more complicated; the nonlinearity and the damping together affects the system response so it is not always obvious what part the damping plays in the response. Zavodney and Shihada[5] investigated the influence of linear viscous damping on the fundamental and principal parametric resonances of SDOF systems possessing quadratic and cubic nonlinearities. The results showed that linear viscous damping plays a significantly different role than in the externally excited oscillator; if reduction in amplitude at resonance is the objective, then in some cases certain critical levels of damping must be exceeded--otherwise one can increase the damping by an order of magnitude but realize less than 5% reduction in the response amplitude.

When the structure possesses more than one mode, as is usually the case, there is the possibility of nonlinear modal coupling. When modes are coupled, it is possible to exchange energy from a directly excited mode to another mode. The end result is that more than one mode is participating in the response and, hence, the structure is vibrating at other frequencies in addition to the frequency of the excited mode--which is not always the same as the excitation frequency. As a result the analysis becomes even more complicated. Attempting to find appropriate mathematical models for these behaviors is extremely difficult. One particular mathematical model may describe one type of behavior, but when the excitation frequency is changed by only 0.1 Hz, the behavior is something else--qualitatively and quantitatively--i.e., a bifurcation has occurred. The nonlinear coupling responsible for this behavior is further enhanced if the structure possesses an internal resonance; an internal resonance occurs whenever any natural frequencies are commensurate (i.e., in a 2:1, 3:1, 3:2, etc. ratio). When a structure simultaneously possesses an internal resonance and appropriate nonlinear coupling

terms, it is possible for one mode to parametrically excite another mode; this phenomenon is called an autoparametric resonance. When this happens, the nonlinear effects are greatly intensified and completely dominate the response. Nayfeh and Zavodney[6] and Balachandran and Nayfeh[7] provided a theoretical model and experimental results showing that such behavior can lead to long-time responses that are not steady-state; mathematically a Hopf bifurcation has occurred.

The objective of this project was to conduct experiments on nonlinear SDOF and MDOF flexible structures to study the effects of damping on parametric and autoparametric resonance. The experiments were performed on structures fabricated from prismatic beams and lumped masses. These types of structures were easy to prepare and tune--i.e., by adjusting the position of the masses the resonant frequencies could be changed. This was essential for the MDOF structure when an internal resonance was desired.

3.0 SINGLE-DEGREE-OF-FREEDOM STRUCTURE

3.1 Theoretical Analysis

Many structural elements can be modelled as slender beams and concentrated masses. When the support undergoes motion, the beam is subject to vibration--either external or parametric, or both. In this section the governing equations are derived for inplane flexural vibration of a thin elastic prismatic cantilever beam subject to parametric vibration at the base. The nonlinear terms arising from the curvature and coupling effects are retained. Galerkin's method is used to discretize the governing nonlinear partial differential equation of motion. The linear eigenvalue problem is solved to determine the eigenvalues, and the eigenfunction is used to determine the coefficients of the time modulation equation. A multiple-scales perturbation solution is obtained for the temporal modulation equation.

The governing equation of motion is derived using Euler-Bernoulli beam theory. The beam, shown in Figure 1, is cantilevered at the oscillating support, has a length L, and carries a concentrated mass m at an arbitrary distance s = d along the neutral axis of the beam. We assume that the thickness of the beam is so small compared with the length that the effects of shearing deformation and rotatory inertia of the beam can be neglected. Since we are investigating parametric resonances (transverse vibration), we will not consider axial resonances of the beam since the frequency of excitation will be far below the first axial resonance. If the beam is kept relatively short (< 30 beam widths), the transverse vibration is purely in plane (if the lumped mass is symmetrical with the centerline); if the excitation frequency is far below the first torsional mode, then we can neglect the torsional modes of the beam in the analysis. These assumptions are consistent with observations in the laboratory. Also, we do not observe any combination or internal resonances. When the mass is removed and the length is increased by an order of magnitude, combination resonances do occur.

According the Euler-Bernoulli beam theory, the bending moment at any cross-section s is given by

$$M(s) = + EI\,\kappa(s) \quad , \tag{1}$$

where E is the elastic (Young's) modulus, I is the cross-sectional moment of inertia, and κ^{-1} is the radius of curvature at section s. From Figure 1

$$\kappa = \frac{\partial \phi}{\partial s} = \phi_s \tag{2}$$

and

$$\sin\phi = v_s \, , \tag{3}$$

where $\phi(s)$ is the slope. The subscript s denotes a partial derivative with respect to s, and the subscript t or the overdot denotes a partial derivative with respect to time t. Differentiating (3) and substituting into (4), and expanding the radical, we obtain

$$M(s) = EI\, v_{ss} \left[1 + \frac{1}{2} v_s^2 + \frac{3}{8} v_s^4 + \ldots \right] . \tag{4}$$

For the analysis presented here, up to third-order terms will be retained.

The moment in the beam is the result of three sources:

$$M(s) = M_1 + M_2 + M_3 \tag{5}$$

where M_1 is the moment at s due to the lateral inertia of the beam element $d\xi$ and the mass m, M_2 is the moment due to the longitudinal inertia of the beam element $d\xi$ and the mass m, and M_3 is the moment at s caused by the angular acceleration of the mass m due to its mass moment of inertia J. In the derivation that follows, the following nomenclature will be used:

x,y	-	Newtonian Cartesian reference frame
g	-	acceleration of gravity
s	-	reference variable along beam
ξ	-	variable of integration along beam
$d\xi$	-	differential length of beam element
L	-	beam length
w	-	beam width
t	-	beam thickness
d	-	position of mass center of mass m
ρ	-	mass density of homogeneous beam per unit length
c	-	coefficient of viscous damping
m	-	mass of concentrated weight on beam
J	-	polar moment of inertia of mass m
$v(\xi,t)$	-	lateral displacement of beam element $d\xi$
$u(\xi,t)$	-	longitudinal displacement of beam element $d\xi$
$\phi(s)$	-	angle with respect to vertical of beam at s
$\kappa(s)$	-	curvature of beam at s
I	-	cross-sectional moment of inertia of beam
E	-	modulus of elasticity of beam

Using the sign convention indicated in Figure 1, the total moment of the beam section from s to L due to the lateral (inertial) motion is

$$- \int_s^L \rho \ddot{v}(\xi,t) \left[\int_s^\xi \cos\phi(\eta,t)\, d\eta \right] d\xi \ . \tag{6}$$

The moment due to m is similarly obtained as

$$- m \ddot{v}(d,t) \left[\int_s^d \cos\phi(\xi)\, d\xi \right], \tag{7}$$

and the moment of the assumed viscous damping is

$$- \int_s^L c \dot{v}(\xi,t) \left[\int_s^\xi \cos\phi(\eta,t)\, d\eta \right] d\xi \ . \tag{8}$$

Combining these and using the Dirac delta function to locate the mass, we obtain

$$M_1 = - \int_s^L \left\{ [\rho + m\, \delta(\xi - d)]\, \ddot{v} + c\, \dot{v} \right\} \left[\int_s^\xi \cos\phi\, d\eta \right] d\xi \ . \tag{9}$$

In a similar fashion, we obtain

$$M_2 = - \int_s^L \left\{ \rho[\ddot{u} - g] + m\, \delta(\xi - d)\, [\ddot{u} - g] \right\} \left[\int_s^\xi \sin\phi\, d\eta \right] d\xi \tag{10}$$

and

$$M_3 = - J \ddot{\phi}(d,t) = - \int_s^L J\, \delta(\xi - d)\, \ddot{\phi}\, d\xi \ . \tag{11}$$

The longitudinal displacement caused by the shortening effect is

$$u(\xi,t) = \xi - \int_o^\xi \cos\phi(\eta,t)\, d\eta \ , \tag{12}$$

which, when combined with the displacement z(t) of the base yields the total axial displacement

$$u(\xi,t) = \xi - \int_0^{\xi} \cos\phi(\eta,t)\, d\eta + z(t) \ . \tag{13}$$

To obtain the governing differential equation, we differentiate (4) and (5) twice with respect to s. This involves differentiating (9), (10), and (11) and using Leibnitz' rule. The second time derivative of the axial displacement $u(\xi,t)$ is obtained by substituting for $\cos\phi$ from (3) and differentiating (13) with respect to t twice to yield

$$\ddot{u} = \frac{1}{2} \int_0^{\xi} (v_\eta^2)_{tt}\, d\eta + \ddot{z}(t) + \ldots \tag{14}$$

Differentiating (9)-(11) twice with respect to s and using (14) yields

$$\frac{\partial^2 M_1}{\partial s^2} = -\left(1 - \frac{1}{2} v_s^2\right) \left\{ [\rho + m\,\delta(s-d)]\, \ddot{v}(s,t) + c\,\dot{v}(s,t) \right\}$$

$$- v_s v_{ss} \int_s^L \left\{ [\rho + m\,\delta(\xi - d)]\, \ddot{v}(\xi,t) + c\,\dot{v}(\xi,t) \right\} d\xi \ , \tag{15}$$

$$\frac{\partial M_2}{\partial s} = v_s \left\{ \rho \int_s^L \left[\frac{1}{2} \int_0^{\xi} \left(v_\eta^2\right)_{tt} d\eta - \ddot{z} \right] d\xi + \rho g (L - s) \right.$$

$$+ m \int_s^L \delta(\xi - d) \left[\frac{1}{2} \int_0^{\xi} \left(v_\eta^2\right)_{tt} d\eta + \ddot{z} - g \right] d\xi$$

$$= v_s N \ , \tag{16}$$

where

$$N = + \frac{1}{2} \rho \int_s^L \left[\int_0^{\xi} \left(v_\eta^2\right)_{tt} d\eta \right] d\xi - \frac{1}{2} m \int_s^L \delta(\xi - d) \left[\int_0^{\xi} \left(v_\eta^2\right)_{tt} d\eta \right] d\xi$$

$$+ m(\ddot{z} - g) \int_s^L \delta(\xi - d)\, d\xi + \rho L \left(1 - \frac{s}{L}\right)(\ddot{z} - g) \quad , \tag{17}$$

and

$$\frac{\partial^2 M_2}{\partial s^2} = (N v_s)_s \quad , \tag{18}$$

and

$$\frac{\partial^2 M_3}{\partial s^2} = + \frac{\partial}{\partial s} \left\{ J\, \delta(s - d) \left[\ddot{v}_s \left(1 + \frac{1}{2} v_s^2\right) + v_s \dot{v}_s^2 \right] \right\} \quad . \tag{19}$$

Differentiating (4) and (5) twice with respect to s and substituting for the $\partial^2 M_n / \partial s^2$ from (15), (18) and (19) yields the governing equation

$$EI \left(v_{ssss} + \frac{1}{2} v_{ssss} v_s^2 + 3 v_s v_{ss} v_{sss} + v_{ss}^3 \right) + \left(1 - \frac{1}{2} v_s^2 - \ldots\right)$$

$$\cdot\, [\rho + m\, \delta(s - d)]\, \ddot{v} - \frac{\partial}{\partial s}(N v_s) + v_s v_{ss} \int_s^L [\rho + m\, \delta(\xi - d)]\, \ddot{v}\, d\xi$$

$$- \frac{\partial}{\partial s} \left\{ J\, \delta(s - d) \left[\ddot{v}_s \left(1 + \frac{1}{2} v_s^2 + \ldots\right) + v_s \dot{v}_s^2 \right] \right\}$$

$$+ \left(1 - \frac{1}{2} v_s^2 - \ldots\right) c \dot{v} + v_s v_{ss} \int_s^L c\, \dot{v}\, d\xi = 0 \, . \tag{20}$$

This field equation is subject to the following boundary conditions:

$$v(0,t) = 0 \quad , \tag{21}$$

$$v_s(0,t) = 0 \quad , \tag{22}$$

$$v_{ss}(L,t) = 0 \quad , \tag{23}$$

$$v_{sss}(L,t) = 0 \quad . \tag{24}$$

The governing problem (20)-(24) is nonlinear and does not admit a closed-form solution. Therefore, an approximate solution was sought that satisfies both the equation and the boundary conditions. Since the boundary conditions are spatial and independent of time, the solution of the nonlinear problem is assumed to take the form

$$v(s,t) = \sum_n r\,\psi_n(s)\,G_n(t) \tag{25}$$

where r is a scaling factor, $\psi_n(s)$ is the shape function of the nth linear mode, and $G_n(t)$ is the time modulation of the nth mode.

The undamped linear free vibration problem is governed by

$$EIv^{iv} + [\rho + m\,\delta(s-d)]\,\ddot{v} = 0 \tag{26}$$

and is subject to the same boundary conditions given by (21)-(24). Without loss of generality, we will solve explicitly for the first mode, with the understanding that the eigenfunction of the nth mode, and its associated eigenvalue, correspond to the nth characteristic.

To solve this problem, the beam is separated into two parts at the concentrated mass. The appropriate boundary conditions are imposed on the two sections to determine the coefficients of the general solution.

The general solution can be stated as the composite function

$$\begin{aligned}\psi(s) = {} & C_1\left[\left(\sin\frac{k}{L}s - \sinh\frac{k}{L}s\right) - \Lambda\left(\cos\frac{k}{L}s - \cosh\frac{k}{L}s\right)\right] \\ & + C_1\,U(s-d)\left\{h_1\left[\sin\frac{k}{L}(s-d) - \sinh\frac{k}{L}(s-d)\right]\right. \\ & + h_3\left[\cos\frac{k}{L}(s-d) - \cosh\frac{k}{L}(s-d)\right] \\ & - \Lambda h_2\left[\sin\frac{k}{L}(s-d) - \sinh\frac{k}{L}(s-d)\right] \\ & \left. - \Lambda h_4\left[\cos\frac{k}{L}(s-d) - \cosh\frac{k}{L}(s-d)\right]\right\}\ ,\end{aligned} \tag{27}$$

where C_1 is an arbitrary constant and k is the characteristic root of the frequency equation

$$\frac{4\rho^2L^4}{mJ}\,[h_1 h_4 - h_2 h_3]$$

$$+ \left(\frac{2\rho Lk^3}{m}\right)\left[h_1\left(\sin\frac{kd}{L} + \sinh\frac{kd}{L}\right) + h_2\left(\cos\frac{kd}{L} - \cosh\frac{kd}{L}\right)\right]$$

$$+ \left(\frac{2\rho L^3 k}{J}\right) \left[h_4 \left(\sin \frac{kd}{L} - \sinh \frac{kd}{L}\right) - h_3 \left(\cos \frac{kd}{L} - \cosh \frac{kd}{L}\right)\right]$$

$$+ 2k^4 \left(1 - \cos \frac{kd}{L} \cosh \frac{kd}{L}\right) = 0 \quad . \tag{28}$$

The frequency of oscillation is given by

$$\omega^2 = \frac{EI}{\rho} \left(\frac{k}{L}\right)^4 \quad . \tag{29}$$

Since a finite number of modes is being considered, this continuous system can be discretized by any of the variational methods available such as Rayleigh-Ritz. When the assumed comparison function is the eigenfunction in particular, the procedure is known as Galerkin's method. It requires multiplying each term by the eigenfunction $\psi(s)$ and then integrating from 0 to L. When this procedure is applied to (20) we obtain

$$\ddot{G} + 2\varepsilon\hat{\zeta}\dot{G} + \theta^2 (1 - \varepsilon f \cos\Omega t)G + \varepsilon\hat{\alpha}G^3 + \varepsilon\kappa_1 G\dot{G}^2$$
$$+ \varepsilon\kappa_2 G^2\ddot{G} - 2\varepsilon\hat{\zeta}\, v\, G^2\dot{G} = 0 \quad , \tag{30}$$

where

$$\hat{\zeta} = \frac{cH_{21}}{2\varepsilon\rho[H_{11} + \mu H_{12} + j\lambda^2 H_{13}]} \quad ,$$

$$\theta^2 = \frac{\frac{EI}{\rho}\left(\frac{k}{L}\right)^4 [H_{31} + \mu H_{32}] - \frac{g}{L}(H_{33} + \mu H_{34})}{[H_{11} + \mu H_{12} + j\lambda^2 H]} \quad ,$$

$$f = \frac{\Omega^2 \Gamma [H_{33} + \mu H_{34}]}{\varepsilon\theta^2 L[H_{11} + \mu H_{12} + j\lambda^2 H_{13}]} \quad ,$$

$$\hat{\alpha} = \frac{EI\lambda^2}{\varepsilon\rho L^4} \frac{[k^4(H_{41} + \mu H_{42}) + H_{43}]}{[H_{11} + \mu H_{12} + j\lambda^2 H_{13}]} ,$$

$$\kappa_1 = \frac{[H_{51} + \mu H_{52} + \lambda^2 jH_{53}]\lambda^2}{\varepsilon[H_{11} + \mu H_{12} + j\lambda^2 H_{13}]} ,$$

$$\kappa_2 = \frac{[H_{61} - H_{62} + \mu H_{63} - \mu H_{64} + j\lambda^2 H_{65}]\lambda^2}{\varepsilon[H_{11} + \mu H_{12} + j\lambda^2 H_{13}]} ,$$

$$\nu = \lambda^2 \frac{H_{71}}{H_{21}} . \tag{31}$$

The H_{ij} are defined in Appendix C.

A new time scale is introduced as

$$\tau = \theta t . \tag{32}$$

The time derivatives become

$$\frac{d}{dt} = \theta \frac{d}{d\tau} \quad \text{and} \quad \frac{d^2}{dt^2} = \theta^2 \frac{d^2}{d\tau^2} . \tag{33}$$

Hence, equation (30) is transformed into

$$G_{\tau\tau} + 2\varepsilon\zeta G_\tau + [1 - \varepsilon f \cos(\phi\tau)]G + \varepsilon\alpha G^3 + \varepsilon\kappa_1 G G_\tau^2$$

$$+ \varepsilon\kappa_2 G^2 G_{\tau\tau} - 2\varepsilon\zeta\nu G^2 G_\tau = 0 , \tag{34}$$

where

$$\zeta = \frac{\hat{\zeta}}{\theta} , \quad \alpha = \frac{\hat{\alpha}}{\theta} , \quad \text{and} \quad f = \frac{\Omega}{\theta} . \tag{35}$$

This equation contains cubic nonlinearities and nonlinear damping, and hence it does not lend itself to a closed-form solution. It must be analyzed by a perturbation or a numerical technique. Moreover, the coefficients are evaluated according to (31) and (C1) and are dependent on $\psi(s)$.

When the amplitude of vibration gets large, $\psi(s)$ may no longer approximate the mode shape, and consequently the coefficients of the time modulation equation have increasing error.

A first-order uniform solution of (34) is sought using the method of multiple scales in the form

$$G(\tau;\varepsilon) = u_0(T_0,T_1) + \varepsilon u_1(T_0,T_1) + \ldots \tag{36}$$

The detuning σ was introduced into the normalized natural frequency of the system as

$$1 = \left(\tfrac{1}{2}\phi\right)^2 + \varepsilon\sigma \tag{37}$$

because the frequency of the response for a principal parametric resonance is exactly one half that of the excitation. Following the standard procedure of eliminating secular terms, we obtain

$$G(\tau) = a\cos\left(\tfrac{1}{2}\phi\tau + \beta\right) + \varepsilon a\left\{\frac{f}{4\phi^2} + \frac{a^2}{32}\left(\frac{4\alpha}{\phi^2} - \kappa_1 - \kappa_2\right)\cos\left[3\left(\tfrac{1}{2}\phi\tau + \beta\right)\right]\right.$$

$$\left. + \frac{\zeta\nu}{8\phi}a^2\sin\left[3\left(\tfrac{1}{2}\phi\tau + \beta\right)\right]\right\} + \ldots, \tag{38}$$

where a and β are given by

$$\dot{a} = -\varepsilon\zeta\left(\tfrac{1}{4}\nu a^2\right)a + \frac{\varepsilon f a}{2\phi}\sin 2\beta\ , \tag{39}$$

$$a\dot{\beta} = -\frac{\varepsilon}{\phi}\left(\sigma + \alpha_e a^2\right)a + \frac{\varepsilon f a}{2\phi}\cos 2\beta\ . \tag{40}$$

Periodic solutions of (34) correspond to the fixed points (i.e., constant solutions) of the modulation equations (39) and (40), which in turn correspond to $\dot{a} = 0$ and $\dot{\beta} = 0$. By squaring and summing (39) and (40) and after imposing $\dot{a} = 0$ and $\dot{\beta} = 0$, we obtain

$$a^2 = 0\ ,\ \frac{1}{2q}\left[-r \pm \sqrt{r^2 - 4qs}\ \right]\ , \tag{41}$$

where

$$q = \frac{1}{4}(1 - \varepsilon\sigma)\zeta^2 v^2 + \alpha_\theta^2 \ ,$$

$$r = 2\sigma(\alpha_\theta + \varepsilon\zeta^2 v) - 2v\zeta^2 \ ,$$

$$s = 4(1 - \varepsilon\sigma)\zeta^2 + \sigma^2 - \frac{1}{4}f^2 \ ,$$

$$\alpha_\theta = \frac{1}{4}[3\alpha + (1 - \varepsilon\sigma)(\kappa_1 - 3\kappa_2)] \ , \tag{42}$$

where ϕ^2 was replaced with $4(1 - \varepsilon\sigma)$ from (37). A trivial fixed point is unstable if and only if

$$f > f_{crit} = 2\sqrt{\sigma^2 - \sigma^2\zeta^2} \ , \tag{43}$$

otherwise it is stable.

A nontrivial fixed point is stable if and only if the real parts of both eigenvalues of the coefficient matrix in (44) and (45) are less than or equal to zero:

$$\dot{a}_1 = \varepsilon\left[-\zeta + \frac{3}{4}\zeta v a_o^2 + \frac{f}{2\phi}\sin 2\beta_o\right]a_1 + \left[\frac{\varepsilon f}{\phi}a_o\cos 2\beta_o\right]\beta_1 \ , \tag{44}$$

$$\dot{\beta}_1 = \left[\frac{2\varepsilon\alpha_\theta a_o}{\phi}\right]a_1 + \left[\frac{-\varepsilon f}{\phi}\sin 2\beta_o\right]\beta_1 \ . \tag{45}$$

3.2 Experimental Analysis

Experiments were performed on the structure shown in Figure 1. The excitation was a base displacement (along the axis in the vertical direction) at a frequency nearly twice that of the first flexural mode of the cantilever beam. The resonant response that ensues is called a principal parametric resonance. The experiments consisted of frequency sweeps at constant-amplitude acceleration and amplitude sweeps at constant frequency. The amplitude was held constant by a computer-controlled feedback loop; as the frequency was changed, the corresponding excitation amplitude required adjustment to keep the acceleration level constant. The excitation level was chosen as large as possible without causing excessive amplitudes of displacement; a maximum level of 0.350 g's was selected. The response was measured by strain gages mounted on the beam.

The first beam, called beam 1, was a very flexible steel beam. Its dimensions and properties are listed in Table 1 in Appendix A. Using these numbers, a theoretical frequency-response curve was obtained using the theory, and is shown in Figure 2. This figure shows hardening behavior because the curve bends over to the right. The theoretical amplitude response for this beam is shown in Figure 3 for a direct excitation to the principal parametric resonance. It shows nonlinear behavior because the amplitude curve bends over. If the system were linear, the line would be perfectly vertical.

The time history of the table acceleration and the response displacement (strain gage) is shown in Figure 4. From this figure it is easy to see that the response is occurring at exactly one-half the frequency of the excitation. It also shows that the excitation amplitude is constant throughout the experiment. The frequency response for beam 1 for two levels of excitation is shown in Figure 5. The amplitude response for beam 1 is shown in Figure 6 for an excitation frequency of 2.000. Both of these figures are in agreement with the theory.

The flexibility of beam 1 was such that it allowed large amplitudes and hence, suffered fatigue cracking; it failed prematurely (meaning that a complete battery of experiments could not be performed on it before it failed). Therefore, a thicker and longer and wider beam was selected for the damping treatment experiments. Its dimensions and properties are also shown in Table 1 in Appendix A.

The theoretical frequency response for beam 2 for two excitation levels is shown in Figure 7. These curves are quite similar to the ones obtained for beam 1. The experimental frequency response of the undamped structure (here undamped means before any damping treatment was applied) for two levels of excitation is shown in Figure 8. It shows that the system is softening because it bends to the left. This behavior was neither expected nor predicted. These experiments were conducted by increasing the frequency of excitation very slowly while simultaneously keeping the table acceleration constant. The arrows indicate jumps; for example, when the excitation level is 0.350 g's (denoted by the circles), the response jumps up to the large amplitude and then slowly decreases as the frequency is increased. When the frequency is

decreased, the response follows the same curve and extends it into an overhang. The response eventually jumps down to the trivial response. The response for the 0.250-g excitation level is qualitatively the same but at a lower amplitude.

Additional experiments at the same levels of excitation were performed with one and two strips of viscoelastic damping treatment applied, as shown in Figures 9 and 10. For the 0.250-g excitation level, the principal parametric resonance was completely suppressed with only two strips of damping treatment applied. Two 0.10-inch strips of damping treatment covers only 5% of the surface area of beam 2 between the support and the lumped mass.

The regions of parametric resonance in the excitation-amplitude versus the excitation-frequency plane were also determined. These experiments were performed by repeating the frequency response experiments for different levels of excitation and plotting only the bifurcation points where either the trivial response becomes unstable (during both a sweep up and a sweep down) or a nontrivial response jumps to the trivial response. An example is shown in Figure 11. This figure shows three boundaries which divide the domain into two regions; the inner-most region (bounded by the center curve and the right curve) represents the loss of trivial stability for the linear system. The additional curve on the far left represents the extension of the instability region caused by the overhang due to the softening nonlinearity in the system. Because nontrivial responses exist at excitation frequencies below those predicted by linear theory, the system is said to possess a subcritical instability. The amount of overhang is reduced for increased damping levels, as shown in Figures 12 and 13.

A summary of the bifurcation boundaries showing only the region where nontrivial responses exist is shown in Figure 14. As the equivalent linear viscous damping coefficient increases, the instability regions move away from the frequency axis. The bifurcation boundaries predicted from the theory using the experimentally measured damping coefficients are shown in Figure 16. The values of the equivalent damping coefficient were obtained from the free response data shown in Figure 15. The Eigensystem Realization Algorithm (ERA) was used to

estimate the damping coefficients and natural frequencies. The theoretical curves are compared with the corresponding experimental data in Figures 17, 18, and 19. The agreement is excellent.

3.3 Discussion of Results

Principal parametric resonances can be attenuated by the application of constrained-layer damping treatment. The experiments show that just a small quantity of the material can be effective. The theoretical model did a good job in predicting the stability boundaries for the parametric resonance, but was unable to predict the softening behavior of beam 2. It also did not predict the jump down. More sophisticated modelling to account for damping and higher order perturbation analysis may be required.

In summary, increased application of the damping treatment reduces the amplitude of the resonant response, and reduces the region of parametric instability. The only way to completely suppress a parametric resonance is to move the instability region far enough away from the axis so that it is completely removed from the area of interest.

4.0 MULTIPLE-DEGREE-OF-FREEDOM STRUCTURE

4.1 Theoretical Analysis

Several researchers have investigated this model and have developed mathematical models. Following Haddow et al[10], the equations describing the displacement of m_1 and m_2 are given by

$$v_1(x,t) = \varepsilon l_1 \phi_{11}(x) u_1(\tau) + \varepsilon l_1 \phi_{12}(x) u_2(\tau) \tag{46}$$

$$v_2(y,t) = \varepsilon l_1 \phi_{21}(y) u_1(\tau) + \varepsilon l_1 \phi_{22}(y) u_2(\tau) \tag{47}$$

where the u_i are given by

$$\ddot{u}_1 + 2\varepsilon\mu_1\dot{u}_1 + \omega_1^2 u_1 + \varepsilon X_{11}(\dot{u}_1)^2 + \varepsilon X_{12}\dot{u}_1\dot{u}_2$$

$$+ \varepsilon X_{13}(\dot{u}_2)^2 + \varepsilon Y_{11} u_1 \ddot{u}_1 + \varepsilon Y_{12} u_1 \ddot{u}_2$$

$$+ \varepsilon Y_{13} u_2 \ddot{u}_1 + \varepsilon Y_{14} u_2 \ddot{u}_2$$

$$+ 2\varepsilon Z_{11} F \Omega u_1 \cos\Omega\tau + 2\varepsilon Z_{12} F \Omega u_2 \cos\Omega\tau$$

$$= 2 F K_1 \cos\Omega\tau \tag{48}$$

$$\ddot{u}_2 + 2\varepsilon\mu_2\dot{u}_2 + \omega_2^2 u_2 + \varepsilon X_{21}(\dot{u}_1)^2 + \varepsilon X_{22}\dot{u}_1\dot{u}_2$$

$$+ \varepsilon X_{23}(\dot{u}_2)^2 + \varepsilon Y_{21} u_1 \ddot{u}_1 + \varepsilon Y_{22} u_1 \ddot{u}_2$$

$$+ \varepsilon Y_{23} u_2 \ddot{u}_1 + \varepsilon Y_{24} u_2 \ddot{u}_2$$

$$+ 2\varepsilon Z_{21} F \Omega u_1 \cos\Omega\tau + 2\varepsilon Z_{22} F \Omega u_2 \cos\Omega\tau$$

$$= 2 F K_2 \cos\Omega\tau \tag{49}$$

The constants X_{ij}, Y_{ij}, and Z_{ij} are defined in reference [10]. The two cases of particular interest are when $\Omega \approx \omega_1$ and $\Omega \approx \omega_2$ which corresponds to a direct excitation to the first mode and to the second mode.

First Mode Excitation

For the case of an internal resonance and excitation to the first mode, the detuning parameters are defined as

$$\omega_2 = 2\omega_1 + \varepsilon\sigma_1 \quad \text{and} \quad \Omega = \omega_1 + \varepsilon\sigma_2 \ , \tag{50}$$

From a multiple-scales perturbation analysis, the following amplitude-and phase-modulation equations are obtained:

$$a_1' + \mu_1 a_1 - a_1 a_2 \sin\gamma_1 - F \sin\gamma_2 = 0 \quad , \tag{51}$$

$$a_1 \alpha_1' + a_1 a_2 \cos\gamma_1 + F \cos\gamma_2 = 0 \quad , \tag{52}$$

$$a_2' + \mu_2 a_2 + a_1^2 \sin\gamma_1 = 0 \quad , \tag{53}$$

$$a_2 \alpha_2' + a_1^2 \cos\gamma_1 = 0 \quad . \tag{54}$$

F is proportional to the amplitude of excitation and

$$\gamma_1 = \sigma_1 (\varepsilon t) - 2\alpha_1 + \alpha_2 \quad , \tag{55}$$

$$\gamma_2 = \sigma_2 (\varepsilon t) - \alpha_1 \quad . \tag{56}$$

The steady-state solutions correspond to

$$a_1' = a_2' = 0 \quad \text{and} \quad \gamma_1' = \gamma_2' = 0 \quad , \tag{57}$$

and when imposed on equations (51)-(56) yield

$$\alpha_1' = \sigma_2 \tag{58}$$

$$\alpha_2' = 2\sigma_2 - \sigma_1 \tag{59}$$

$$a_2 = a_1^2 \Big/ \left(\mu_2^2 + (2\sigma_2 - \sigma_1)^2\right)^{1/2} \tag{60}$$

$$a_1^6 + 2[\mu_1 \mu_2 - \sigma_2 (2\sigma_2 - \sigma_1)]\, a_1^4$$

$$+ \left[\mu_2^2 + (2\sigma_2 - \sigma_1)^2\right] \left[\left(\sigma_2^2 + \mu_1^2\right) a_1^2 - F^2\right] = 0 \, . \tag{61}$$

The stability of these solutions is obtained by perturbing each steady-state solution and studying the behavior of the disturbance. Substituting

$$a_1 = a_{10} + a_{11} \quad , \qquad a_2 = a_{20} + a_{21} \quad , \tag{62}$$

$$\gamma_1 = \gamma_{10} + \gamma_{11} \ , \qquad \gamma_2 = \gamma_{20} + \gamma_{21} \tag{63}$$

into equations (51)-(54) yields

$$\dot{a}_{11} + (\mu_1 - a_{20} \sin\gamma_{10}) a_{11} - a_{10} \sin\gamma_{10} a_{21}$$
$$- a_{10} a_{20} \cos\gamma_{10} \gamma_{10} - F \cos \gamma_{20} \gamma_{22} \ , \tag{64}$$

$$(\sigma_2 + a_{20} \cos\gamma_{10}) a_{11} + a_{10} \cos\gamma_{10} a_{21}$$
$$- a_{10} a_{20} \sin\gamma_{10} \gamma_{11} - a_{10} \dot{\gamma}_{21} - F \sin\gamma_{20} \gamma_{21} \ , \tag{65}$$

$$2a_{10} \sin\gamma_{10} a_{11} + \dot{a}_{21} + \mu_2 a_{21} + a_{10}^2 \cos\gamma_{10} \gamma_{11} = 0 \ , \tag{66}$$

$$2a_{10} \cos\gamma_{10} a_{11} + (2\sigma_2 - \sigma_1) a_{21} + a_{20} \dot{\gamma}_{11} - a_{10}^2 \sin\gamma_{10} \gamma_{11}$$
$$- 2a_{20} \dot{\gamma}_{21} = 0 \ . \tag{67}$$

Since these equations are linear and have constant coefficients, it follows that

$$\dot{a}_{i1} = \lambda a_{i1} \quad \text{and} \quad \dot{\gamma}_{i1} = \lambda \gamma_{i1} \tag{68}$$

where λ is an eigenvalue of the coefficient matrix obtained by substituting (68) into (64)-(67). Any solution of (58)-(61) is stable if the real parts of all eigenvalues are less than zero.

Second Mode Excitation

For the case of internal resonance and an excitation to the second mode, we put

$$\omega_2 = 2\omega_1 + \varepsilon\sigma_1 \quad \text{and} \quad \Omega = \omega_2 + \varepsilon\sigma_2 \tag{69}$$

and obtain the following amplitude- and phase-modulation equations:

$$a_1' + \mu_1 a_1 - a_1 a_2 \sin\gamma_1 = 0 \quad , \tag{70}$$

$$a_1 \alpha_1' + a_1 a_2 \cos\gamma_1 = 0 \quad , \tag{71}$$

$$a_2' + \mu_2 a_2 + a_1^2 \sin\gamma_1 - F \sin\gamma_2 = 0 \quad , \tag{72}$$

$$a_2 \alpha_1' + a_1^2 \cos\gamma_1 + F \cos\gamma_2 = 0 \quad . \tag{73}$$

F is again proportional to the excitation amplitude and

$$\gamma_1 = \sigma_1 (\varepsilon t) - 2\alpha_1 + \alpha_2 \quad , \tag{74}$$

$$\gamma_2 = \sigma_2 (\varepsilon t) - \alpha_2 \quad . \tag{75}$$

The steady-state solutions correspond to

$$a_1' = a_2' = 0 \quad \text{and} \quad \gamma_1' = \gamma_2' = 0 \quad , \tag{76}$$

and, when substituted into (70)-(73) yields

$$a_1 (\mu_1 - a_2 \sin\gamma_1) = 0 \tag{77}$$

$$a_1 \left[\tfrac{1}{2} (\sigma_1 + \sigma_2) + a_2 \cos\gamma_1 \right] = 0 \tag{78}$$

$$\mu_2 a_2 + a_1^2 \sin\gamma_1 - F \sin\gamma_2 = 0 \tag{79}$$

$$\sigma_2 a_2 + a_1^2 \cos\gamma_1 + F \cos\gamma_2 = 0 \tag{80}$$

There are two possible solutions to these equations: the solution which corresponds to the linear problem given by

$$a_1 = 0 \quad \text{and} \quad a_2 = F / \sqrt{\sigma_2^2 + \mu_2^2} \quad , \tag{81}$$

and the nonlinear solution which corresponds to

$$a_1^2 = \frac{1}{2}\left[\sigma_2(\sigma_1 + \sigma_2) - 2\mu_1\mu_2\right] \pm \sqrt{F^2 - \left[\mu_1\sigma_2 + \mu_2\left(\frac{\sigma_1 + \sigma_2}{2}\right)\right]^2} \quad . \tag{82}$$

$$a_2 = \sqrt{\mu_1^2 + \left(\frac{\sigma_1 + \sigma_2}{2}\right)^2} \quad . \tag{83}$$

The stability of the solutions for a direct excitation to the second mode is obtained in a similar fashion. Disturbances are defined in (62) and (63); the governing equations are found to be

$$a_{11}' - a_{10}\sin\gamma_{10}\, a_{21} - a_{10}\, a_{20}\cos\gamma_{10}\,\gamma_{11} = 0 \,, \tag{84}$$

$$a_{10}\cos\gamma_{10}\, a_{21} + \frac{1}{2}a_{10}\,\gamma_{11}' - a_{10}\, a_{20}\sin\gamma_{10}\,\gamma_{11} - \frac{1}{2}a_{10}\,\gamma_{21}' = 0 \,, \tag{85}$$

$$\begin{aligned} 2a_{10}\sin\gamma_{10}\, a_{11} + a_{21}' + \mu 2\, a_{21}' + \mu_2\, a_{21} \\ + a_{10}^2\cos\gamma_{10}\,\gamma_{11} - F\cos\gamma_{20}\,\gamma_{21} = 0 \,, \end{aligned} \tag{86}$$

$$\begin{aligned} 2a_{10}\cos\gamma_{10}\, a_{11} + \sigma_2\, a_{21} - a_{10}^2\sin\gamma_{10}\,\gamma_{11} \\ - a_{20}\,\gamma_{21}' - F\sin\gamma_{20}\,\gamma_{21} = 0 \,. \end{aligned} \tag{87}$$

The behavior of the eigenvalues λ determines the stability of the a_i. The nature of the response predicted from these equations will be discussed later.

4.2 Experimental Analysis

Experiments were performed on a multiple-degree-of-freedom (MDOF) structure. The structure used for these experiments is shown in Figure 20; it has been studied by several researchers[6,7,9,10]. It was chosen because it was easy to fabricate and could be easily tuned (to create an internal resonance) by adjusting the lengths of the beams and position of m_2. The

laboratory setup is shown schematically in Figure 21. The shaker was an Unholtz-Dickie model 200 which has a capacity of 1100-lb force. This insured that there was no feedback distortion to the shaker table from the structure. Data was obtained by strain gages attached to the structure and conditioned by a bridge amplifier. The signals were monitored on a digital oscilloscope and captured by a Hewlett-Packard data acquisition system attached to a PC. The experiments consisted of the following:

a) random excitation to determine resonant frequencies and damping coefficients

b) harmonic excitation at constant frequency and amplitude to determine steady-state response levels

c) harmonic excitation at constant amplitude of acceleration and small changes in the frequency of excitation to effectively cause a slow frequency sweep

d) harmonic excitation at constant frequency and small changes in the amplitude of excitation to effectively cause a slow amplitude sweep

e) impulse response resulting from a sudden release from a deformed position approximating the first mode shape, the second mode shape, and a combination of the two

f) impulse response resulting from an impact to the lower mass.

The nature of the response was quite complicated at times, so a variety of the above techniques was used throughout the experimental phase.

The procedure used for applying damping treatment to the SDOF structure was used for the MDOF structure: thin strips of 0.10-inch width damping treatment were applied in increasing numbers to each beam, increasing the damping of the structure.

The responses of this structure--both to impulse and harmonic excitation--were very complicated. The primary source of the quadratic modal-coupling nonlinearity in this structure was due to the asymmetrical geometry. When an internal resonance existed, the nonlinear coupling was significantly intensified and would dominate the response; when this happened, the

response amplitudes were three to four times larger than the linear response. Internal resonance is a measure of the difference between the frequency of the second mode and two times the frequency of the first mode; the detuning parameter is defined by equation (50) or (69). Two amounts of internal detuning were attempted: perfect detuning and slightly positive.

Because the results of the SDOF experiments showed that a parametric resonance could be suppressed with a small amount of damping treatment, the experiments on the MDOF structure were performed at very high levels of excitation. Studies[6,7,9,10] have shown that under some conditions, nonlinear responses can be achieved with very small levels of excitation (on the order of 20 mili-g's). In the experiments performed here, the excitation levels were on the order of 100 mili-g's--five times that required to solicit a nonlinear response. This large excitation level was used to extend the range for which damping treatment could be applied; in other words, more damping treatment could be applied before the nonlinear motion was expected to be suppressed.

For both the tuned and slightly detuned models, damping treatment was applied in five steps. With each application of damping treatment the model was adjusted (i.e., the length of the lower beam) to return it to the original amount of detuning. This was a painstaking procedure because it sometimes required more than a dozen adjustments in the length of the lower beam just to get the results reported here.

The labels in the subsequent figures correspond to the following amounts of damping treatment: the unmarked or "O" for none, "I" for one strip on each beam, "II" for two strips on each beam (17% of the surface area), "III" for four strips on each beam, "IV" for six strips on each beam (50%), and "V" for 100% coverage.

(Nearly) Perfectly Tuned Structure

The structure shown in Figure 20 was tuned such that the first resonance occurred at 4.012 Hz and the second resonance occurred at 8.040 Hz; this corresponded to an internal

resonance detuning of +0.016 Hz, or +0.20% (of the second resonant frequency). This was as close as practically possible to perfect tuning on a real structure; a tuning closer to 0% could be accomplished at the expense of more painstaking effort (i.e., moving the lower beam into the clamped support in 0.001-inch increments and repeating the experiments until the desired detuning is achieved). After each application of damping treatment, the model was retuned to get nearly 0.0% detuning.

The frequency response function for four of these cases is shown in Figures 22 through 26. The structure was forced with low-level random excitation; the input acceleration was measured by an accelerometer mounted on the base clamp and the response displacement was measured by the strain gages. The roots were estimated by a complex exponential curve fit. The reconstructed frequency-response curve from (the estimated roots) is shown as a dashed line in the figures. The accuracy of the curve fit is verified by the closeness to the experimental data near resonance. The curves do show a trend of increasing damping as evidenced by the increased width of the resonance peaks. The estimates of the natural frequencies and damping coefficients are listed in the figure captions and summarized in Table 2 (Appendix A).

The frequency responses for the structure without treatment and for three of the five cases of damping treatment are shown in Figures 27 and 28. This 3-D perspective plot aids in visualizing the qualitative and quantitative changes caused by the increase in the damping. The frequency axis is shown normalized with respect to the excitation frequency because the addition of the damping treatment caused the frequencies to shift slightly. The response amplitude (displacement of m_1) is represented in units corresponding to the ratio of the lower mass displacement amplitude to the length of the lower beam. This scaling provides a convenient comparison of all of the results for all of the experiments. For example, the peak amplitude of the response of the first mode is approximately 8% of the length of the lower beam.

When the first mode is directly excited with a large-amplitude excitation at a frequency near the first resonance, the response is nonlinear, as shown in Figure 27. Figure 27(a) shows the first mode response and Figure 27(b) shows the second mode response. The response of

the untreated structure shows the most interesting behavior. As the frequency of excitation is increased from 0.90, the amplitude initially grows until the second mode is also excited; this happens at a frequency near 0.97. As the frequency is increased further, the amplitude of the first mode decreases dramatically--almost four times! When a Hopf bifurcation occurs near a frequency of 1.0, no steady-state responses are possible. This region is denoted by dots which represents an "average" amplitude. Actual bounds on the modulation are shown later in the amplitude response curves. As the frequency is increased further, the amplitude of response increases until it jumps down (at the first mode frequency of 1.072) to the linear response amplitude. The linear response consists only of the directly excited mode--no other modes are present. If the frequency is decreased from above, the jump up occurs at a lower frequency (i.e., 1.044) than did the jump down during a sweep up. This indicates an overhang or double-valued steady-state response. Theory shows that these two solution branches are connected by an unstable solution branch; this is shown by the dashed line on the response (note--there are no data points for this dashed line because they cannot be realized in the laboratory). Further decreases in the excitation frequency cause the response to follow the same path as that followed during the sweep up. The frequency response is almost symmetrical; because it is slightly skewed to the right, it indicates that a slight positive detuning of the internal resonance is present.

The frequency response curve for the undamped structure shows four distinct types of motions that are possible when the excitation frequency is near the first mode: (1) the linear solution where only the directly excited mode participates in the response, (2) a region where nonlinear coupling is present and causes two steady-state solutions to co-exist--the linear one and a nonlinear one, (3) only a nonlinear response where the modes achieve one steady-state amplitude, and (4) a region bounded by a Hopf bifurcation where no steady-state solutions exist.

As damping treatment was applied to the structure, response amplitudes were attenuated. Certain trends can be observed in Figure 27. The second application of damping treatment [II] consisted of two strips on each beam (17%). Both response peaks are significantly

attenuated. The peak at a frequency below the resonance is almost completely eliminated while the response peak at the upper frequency is reduced in amplitude sufficiently to eliminate the hysteresis. The fourth application of damping treatment [IV] consisted of six strips (50%) and caused further peak broadening and smoothing, resulting in a frequency response that was so broad that it appeared as if it had a large damping coefficient. The fifth application (100%) almost completely eliminated the modal coupling; however, when one compares the harmonic response (i.e., Figure 27) to the stochastic response (Figure 26), it is obvious that linear-based modal analysis techniques (i.e., random excitation) cannot be used to identify this type of nonlinearity. This figure shows that damping treatment can attenuate the nonlinear hysteretic behavior and eliminate the amplitude of the response.

A direct excitation to the second mode is shown in Figure 28. Similar observations can be made from these curves. The frequency axes have been nondimensionalized with respect to the excitation frequency; hence the second mode has a resonance at 1.0 and the first mode, representing a frequency nearly one-half that of the excitation, has a resonance at 2.0 (representing a subharmonic resonance of order 1/2). The indirectly excited mode (i.e., the first mode) remains trivial until it is strongly (and nonlinearly) coupled to the second mode; when it is, the amplitude of the second mode is drastically attenuated. Furthermore, during a sweep down, there is no jump up in the second mode response--only a jump down--because the lower branch merges with the upper branch; it is not a turning point bifurcation as was the case for a direct excitation to the first mode as seen in Figure 27. The first mode response demonstrates both a jump up and a jump down during a frequency sweep. For the fifth application of damping treatment, the nonlinear coupling between the modes is completely suppressed; the end result is that the system behaves like a moderately damped linear system.

Experiments were also conducted to measure the amplitude response at selected frequencies of excitation; these curves are cuts across a particular frequency response curve shown in Figures 27 and 28. For the case of no damping treatment, a cut at a frequency of 1.004 (first-mode excitation) passes through the modulation region. The amplitude response at

this frequency is shown in Figure 29. As the amplitude of excitation is increased, the system experiences a Hopf bifurcation and begins to modulate; *i.e., energy begins to flow back and forth* between the two modes. As the amplitude of excitation is increased, the "average" amplitude tends to increase and the maximum excursions increase.

At a frequency of 1.067 the first-mode frequency response curve (for no damping treatment) in Figure 27(a) shows a double-value steady-state amplituc An amplitude sweep through this region, as shown in Figure 30, reveals a jump phenomenon. ₩hen starting on the lower branch, increasing the amplitude of excitation causes a jump up to the upper branch; decreasing the amplitude of excitation causes a jump down at a lower value of the amplitude of excitation.

When the maximum damping treatment was applied, there was essentially one response; Figure 31 shows an amplitude sweep through this resonance at a frequency of 0.979. Although the curves are quite tame, they still show nonlinear coupling because the second mode is excited. Even though the excitation is driving the first mode, some of the energy is channeled into the second mode.

Amplitude-excitation experiments were also performed for a direct excitation to the second mode. Figure 32 shows a cut across the 0.998 frequency line (the 1.999 frequency line for the first mode) in Figure 28 for the structure with no damping treatment applied. Although the second mode is directly excited, its response is attenuated; instead, the energy goes into the first mode which responds like a parametrically excited SDOF system. An amplitude sweep at the 1.057 frequency line (2.119 for the first mode) in the region of the overhang is shown in Figure 33. During the experiment, the amplitude of the excitation was increased slowly. At 0.32 g's rms, the linear solution became unstable. However, the divergence growth rate was so slow that it was possible to stay on the unstable branch long enough to locate some of the equilibrium points. These are shown in the figure as isolated circles. However, by waiting long enough, the response went to the steady-state nonlinear solution which is shown as a solid line. During a sweep down, the response followed the upper curve rather than the lower curve. At an excitation

level of 0.075 g's rms, the response jumped down to the linear response on the lower branch and remained there. Note that the overhang region in Figure 28 is qualitatively different than the overhang region in Figure 27; hence, it is not surprising to see a qualitative difference in the double solution region of Figures 30 and 33.

When damping treatment was applied to the structure, the regions of multiple solutions contracted and eventually disappeared. For small levels of excitation the linear solution appeared; this is shown in Figure 34. At this frequency of excitation (0.992 for the first mode, 1.991 for the second mode), when the excitation level exceeded 0.09 g's, the second mode saturated. The energy that was put into the system at the second-mode was channeled into the first mode; essentially the second mode frequency was parametrically exciting the first mode. Hence, the first mode response appears to have a parametric type response.

Slightly Detuned Structure

Experiments for a slightly detuned structure were also performed. When the damping treatment was applied, experiments were performed for both the tuned and detuned case. One way to accomplish the same detuning was to increase the mass of m_2; this was accomplished by adding a small screw to m_2. In this case, the resonances for the undamped structure occurred at 3.750 Hz and 7.750 Hz, which corresponds to a detuning of +0.25 Hz, or +3.2%. The results of these experiments are qualitatively the same, if we had more time we would have performed additional experiments with even more detuning. This amount of detuning would have yielded more significant skewing if a smaller excitation level was used. Increased internal-resonance detuning causes increased skewing of the frequency-response curves and in some cases increases the region where modulated responses occur.

The frequency response functions for low-level random excitation for the five cases are shown in Figures 35 through 40. The trends here are similar to those of the tuned structure. The frequency responses to harmonic excitation are shown in Figures 41 and 42. The features are

qualitatively the same as the tuned structure, except that there is more skewing. Had the detuning been greater, the skewing would have been more pronounced. There is one noticeable difference: in Figure 41(b) the steady-state response amplitudes for the undamped (O) case appear to be discontinuous before and after the Hopf bifurcation. The theoretical response curves will show why this is the case. The qualitative behavior for increased application of damping treatment is the same as that for the tuned case.

Experiments to measure the amplitude response at a fixed frequency were also performed for the slightly detuned structure. These amplitude-response curves represent a cut across a particular frequency-response curve at a fixed frequency (3rd axis representing the amplitude of the excitation). The frequencies were chosen to represent different qualitative regions in the frequency-response curves. For the case of no damping treatment, an amplitude sweep at a frequency of 3.82 Hz is shown in Figure 43; this frequency corresponds to a nondimensional frequency of 0.9941for the first mode shown in Figure 41(a) and 0.4922 for the second mode in Figure 41(b) and is central to the modulation region. The amplitude sweep shows nonlinear coupling immediately and at about 0.05g excitation, a Hopf bifurcation takes place that causes the modulation between the modes. Since the motion was quite large, the experiments were not continued for larger amplitudes of excitation because of fear of the structure failing prematurely. The amplitude response for a frequency of excitation in the overhang region on the right side is shown in Figure 44. The amplitude response for the maximum amount of damping treatment applied to the structure is shown in Figure 45. Even though the entire structure is covered with damping treatment, nonlinear modal coupling still occurs. The response amplitudes are attenuated.

When the second mode is excited, an unusual behavior was observed. The frequency response for a direct excitation to the second mode was shown in Figure 42. When an excitation frequency of 7.45 Hz (1.9388 for the first mode and 0.9600 for the second mode) was used, the response was steady state (i.e., constant modal amplitudes) with modal coupling as shown in

Figure 46. This frequency corresponds to the region just left of the modulation region where modal interaction is quite aggressive.

When the frequency was increased slightly to move into the modulation region, an interesting behavior was observed, as shown in Figure 47. Initially, mode 2 responds and mode 1 is trivial, which is essentially the linear response. When the system bifurcates (i.e., Hopf), the modes begin to interact and exchange energy as previously observed when the first mode was excited. However, as the amplitude of excitation is increased further, there is an amplitude of excitation at which the modal interaction ceases! This may have been possible on the previous cases, but since the amplitudes of response were so large, experiments at the larger amplitudes of excitation were not performed. The nature of the "average" amplitudes (indicated by the dashed line) shown in Figure 47 suggests that the second mode is saturated and the first mode is autoparametrically excited.

If the frequency of excitation is increased further so that we are now to the right of the modulation region and in the overhang region, two possible steady-state responses are possible: the linear and the nonlinear. These responses are shown in Figure 48. The figure shows an isolated data point to the right of the bifurcation. This point is the steady-state response immediately after the excitation amplitude was increased. Theoretical analysis shows that this point is unstable. By waiting long enough (about 10 minutes), the response did go to the stable steady-state amplitude predicted by theory. This experiment shows that patience is required when performing experiments on nonlinear structures because some stable solutions require much more time (than one might) think to be realized in the laboratory.

When damping treatment was applied to the structure, the response was generally attenuated. The trends were shown in the frequency domain in Figure 42. For the maximum amount of damping treatment, the amplitude response is shown in Figure 49. This figure shows the same behavior as discussed previously.

4.3 *Discussion of Results and Comparison with Theory*

The experimental results were compared with the theoretical results using values of the damping coefficients and natural frequencies shown in Table 2 (Appendix A) obtained from the random excitation experiments and complex exponential curve fits. The response of the structure to a direct excitation of the first mode is given by equations (60) and (61); typical results are shown in Figure 50. Two additional theoretical curves were generated for increasing amounts of damping; these are shown in Figures 51 and 52. These curves also capture the qualitative behavior seen in the experimental results.

The steady-state response for a direct excitation to the second mode is predicted by equations (77)-(80). Curves were generated for the response to a direct excitation to the second mode. The case of no damping treatment "O" is shown in Figure 53, case "III" is shown in Figure 54, and case "V" is shown in Figure 55. Again, the qualitative features are captured by the theory, including the nature of the bifurcations. Apparently the experiment had more detuning than was measured because the theoretical curves are more symmetrical than the experimental curves.

For the tuned structure, the results are qualitatively the same. Typical results for a direct excitation to the first and second mode are shown in Figures 56 and 57, respectively.

5.0 SUMMARY and CONCLUSIONS

Analysis and experiments were performed to study the effect of commercially available viscoelastic damping treatment on parametric and autoparametric resonances in nonlinear systems. Both SDOF and MDOF structures were used; they consisted of flexible beams and masses.

For the SDOF structure, experiments were conducted at several levels of excitation and it was found that the damping treatment was particularly effective in reducing--and even suppressing--the resonance entirely. The first-order theory predicts the region of the loss of trivial stability very well, but failed to predict the softening behavior of beam 2 and the closing (i.e., merginh of the two branches) of the frequency response curves.

For the 2DOF system, the particular case of an internally resonant structure was considered. The qualitative and quantitative effects of the damping treatment were determined for very large excitation levels; it was observed that increased amounts of damping treatment reduces the nonlinear modal coupling and modulation regions and reduces the amplitude of response. For the case of direct excitation to the second mode, a large amount of damping treatment was capable of eliminating the nonlinear coupling.

Further research should investigate flat plates subject to parametric excitation. The formulation for the problem would be similar to that of the SDOF beam, but allow multi-modal interaction similar to the 2DOF model.

6.0 REFERENCES

1. Zavodney, L.D., "A Theoretical and Experimental Investigation of Parametrically Excited Nonlinear Mechanical Systems," Ph.D. Dissertation, Virginia Polytechnic Institute and State University, 1987.

2. Zavodney, L.D. and Nayfeh, A.H. "The Response of a Single-Degree-of-Freedom System with Quadratic and Cubic Nonlinearities to a Fundamental Parametric Resonance," *Journal of Sound and Vibration*, 1988, Vol 120, 63-93.

3. Zavodney, L.D., Nayfeh, A.H. and Sanchez, N.E., "The Response of a Single-Degree-of-Freedom System with Quadratic and Cubic Nonlinearities to a Principal Parametric Resonance," *Journal of Sound and Vibration*, 1989, Vol 129, 417-442.

4. Zavodney, L.D., Nayfeh, A.H. and Sanchez, N.E., "Bifurcations and Chaos in Parametrically Excited Single-Degree-of-Freedom Systems," *Nonlinear Dynamics*, 1990, Vol 1, 1-21.

5. Zavodney, L.D. and Shihada, S.M., "The Role of Damping in the Suppression of Parametric Resonances in Nonlinear Systems," *Proceedings of Damping '89*, West Palm Beach, Florida, February 8-10, 1989, FBD-1 - FBD-22.

6. Nayfeh, A.H. and Zavodney, L.D., "Experimental Observation of Amplitude- and Phase-Modulated Responses of Two Internally Coupled Oscillators to a Harmonic Excitation," *Journal of Applied Mechanics*, 1988, Vol 110, 706-710.

7. Balachandran, B. and Nayfeh, A.H., "Nonlinear Motions of Beam-Mass Structure," *Nonlinear Dynamics*, 1990, Vol 1, 39-62.

8. Zavodney, L.D. and Nayfeh, A.H., "The Non-Linear Response of a Slender Beam Carrying a Lumped Mass to a Principal Parametric Excitation: Theory and Experiment," *International Journal of Nonlinear Mechanics*, 1989, Vol 24, 105-125.

9. Zavodney, L.D. and Hollkamp, J.J., "Experimental Identification of Internally Resonant Nonlinear Structures Possessing Quadratic Nonlinearity," *Proceedings of the 32nd Structures, Structural Dynamics and Materials Conference*, Baltimore, Maryland, 8-10 April, 1991, 2755-2765.

10. Haddow, A.G., Barr, A.D.S., and Mook, D.T., "Theoretical and Experimental Study of Modal Interaction in a Two-Degree-of-Freedom Structure," *Journal of Sound and Vibration*, 1984, Vol 97, 451-473.

APPENDIX A

TABLE 1. Dimensions and properties of SDOF beam and lumped mass.

Beam	1	2
Length (mm)	125.40	360.36
Width (mm)	15.85	19.16
Thickness (mm)	0.559	1.206
I (mm^4)	6.599×10^{-2}	2.805
E (N/mm)	0.209×10^{6}	0.202×10^{6}
r (gm/mm)	7.762×10^{-3}	0.1798
n (sec^{-1})	1.469×10^{-3}	1.207×10^{-3}
Lumped Mass		
Length (mm)	31.75	31.50
Height (mm)	9.53	19.32
Width (mm)	9.05	25.02
Mass (gm)	25.40	60.04
J (gm-mm^2)	1969	6670
d (mm)	76.20	346.1
Natural Freq (Hz)	15.038	3.462

Table 2. Natural frequencies and damping coefficients estimated from complex exponential curve fits to the frequency response data for low-level random excitation of the 2DOF tuned and slightly detuned structure.

Beam	ω_1	ω_2	$\omega_1\zeta_1$	ζ_1	$\omega_2\zeta_2$	ζ_2	$\lvert a_1 \rvert$	$\lvert a_2 \rvert$
Tuned-0	4.012	8.040	0.410	0.016	0.191	0.004	0.012	0.022
Tuned-I	4.091	8.050	0.569	0.022	0.396	0.008	0.006	0.020
Tuned-II	4.062	8.161	0.485	0.019	0.531	0.010	0.010	0.022
Tuned-III	4.131	8.220	0.476	0.018	0.863	0.017	0.011	0.026
Tuned-IV*	4.175	8.283	0.593	0.023	1.135	0.022		
Tuned-V	4.278	8.620	1.425	0.053	1.731	0.032	0.007	0.011
Detuned-O	3.750	7.750	0.183	0.008	0.195	0.004	0.010	0.027
Detuned-I	3.748	7.802	0.289	0.012	0.318	0.007	0.010	0.024
Detuned-II	3.795	7.846	0.387	0.016	0.506	0.010	0.009	0.024
Detuned-III	3.863	7.861	0.507	0.021	0.782	0.016	0.009	0.027
Detuned-IV	3.906	8.186	0.742	0.031	1.162	0.023	0.004	0.012
Detuned-V	4.005	8.324	0.723	0.029	1.560	0.030	0.008	0.023

Frequency (Hz), damping (percent of critical)

* Estimated from sinusoidal-dwell data because random excitation data unavailable.

Appendix B

$$h_1 = (k_{22} l_{11} - k_{12} l_{12}) / D \quad , \tag{B1}$$

$$h_2 = (k_{22} l_{12} - k_{12} l_{22}) / D \quad , \tag{B2}$$

$$h_3 = (k_{11} l_{12} - k_{12} l_{11}) / D \quad , \tag{B3}$$

$$h_4 = (k_{11} l_{22} - k_{12} l_{12}) / D \quad , \tag{B4}$$

$$D = -\ 2\left[1 + \cos \frac{k}{L} (L - d) \cosh \frac{k}{L} (L - d)\right] \quad , \tag{B5}$$

where

$$k_{11} = \sin \frac{k}{L} (L - d) + \sinh \frac{k}{L} (L - d) \quad , \tag{B6}$$

$$k_{12} = \cos \frac{k}{L} (L - d) + \cosh \frac{k}{L} (L - d) \quad , \tag{B7}$$

$$k_{22} = \sin \frac{k}{L} (L - d) + \sinh \frac{k}{L} (L - d) \quad , \tag{B8}$$

$$l_{11} = -\ \sin k - \sinh k \quad , \tag{B9}$$

$$l_{12} = -\ \cos k - \cosh k \quad , \tag{B10}$$

$$l_{22} = \sin k - \sinh k \quad . \tag{B11}$$

$$\Lambda = \frac{m_{11}}{m_{12}} = \frac{m_{21}}{m_{22}} \tag{B12}$$

where

$$m_{11} = \frac{2\rho L h_1}{m} + k\left(\sin\frac{kd}{L} - \sinh\frac{kd}{L}\right) , \tag{B13}$$

$$m_{12} = \frac{2\rho L h_2}{m} + k\left(\cos\frac{kd}{L} - \cosh\frac{kd}{L}\right) , \tag{B14}$$

$$m_{21} = \frac{2\rho L^3 h_3}{J} - k^3\left(\cos\frac{kd}{L} - \cosh\frac{kd}{L}\right) , \tag{B15}$$

$$m_{22} = \frac{2\rho L^3 h_4}{J} + k^3\left(\sin\frac{kd}{L} - \sinh\frac{kd}{L}\right) . \tag{B16}$$

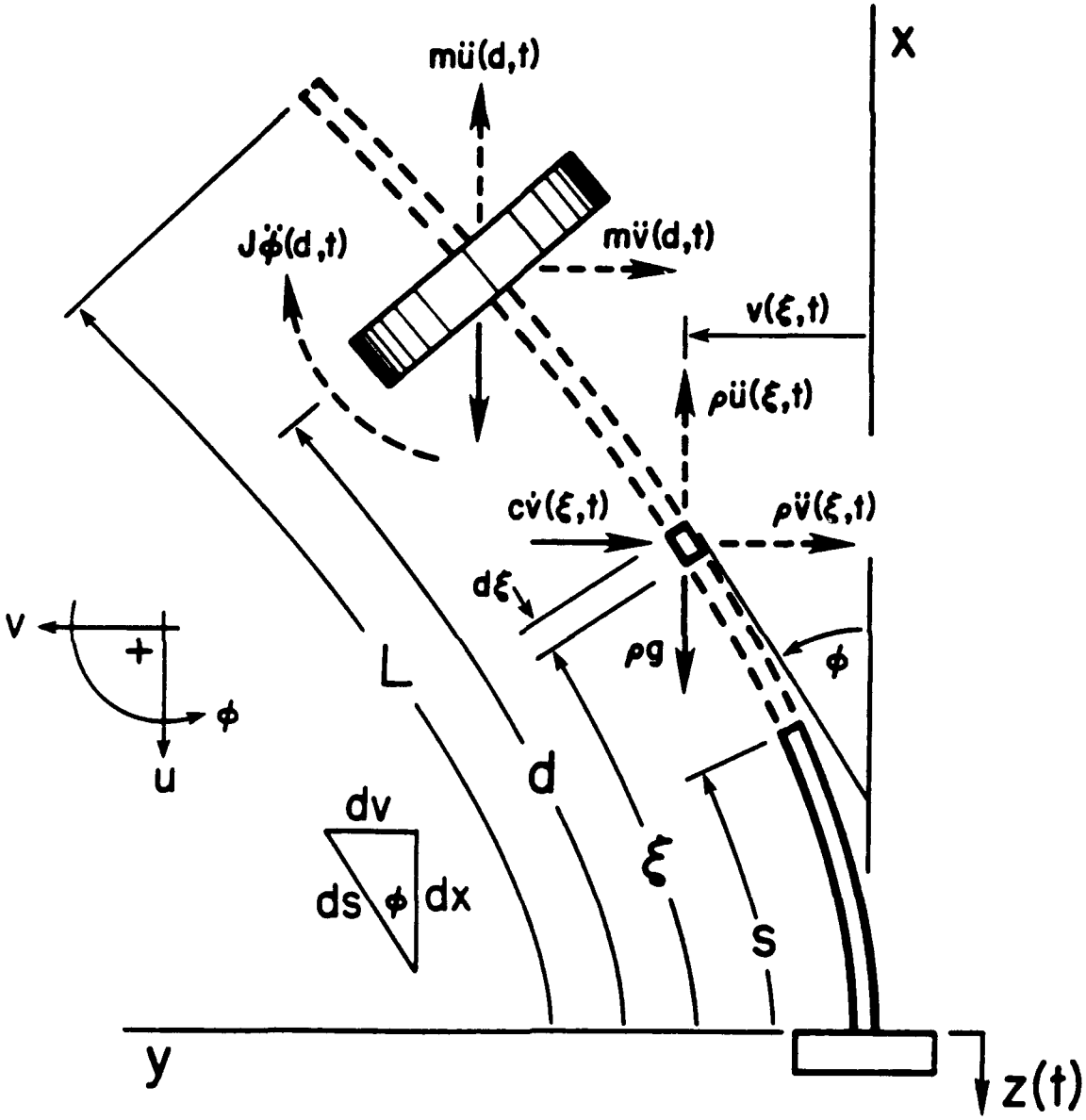


Figure 1. Structure used for the SDOF experiments consisting of a flexible beam carrying a lumped mass. Constrained-layer damping treatment was applied in thin 0.10-inch strips on both sides of the beam to incrementally increase the level of damping in the structure. The base of the beam was clamped to a shaker head that oscillated in the vertical direction.

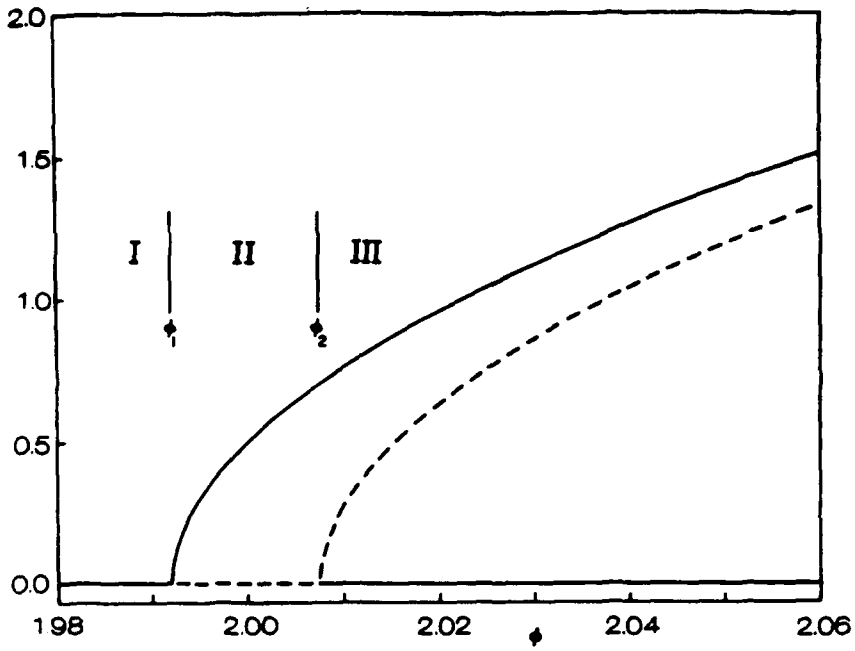


Figure 2. Variation of the steady-state amplitude a* with the nondimensional frequency of excitation of beam 1: (——) stable, (----) unstable, α = 0.0808, κ_1 = 0.4163, κ_2 =0.1716, ζ = 0.1716, v = 0.2577, f = 0.01653 (0.230-g level), ε = 1.00.

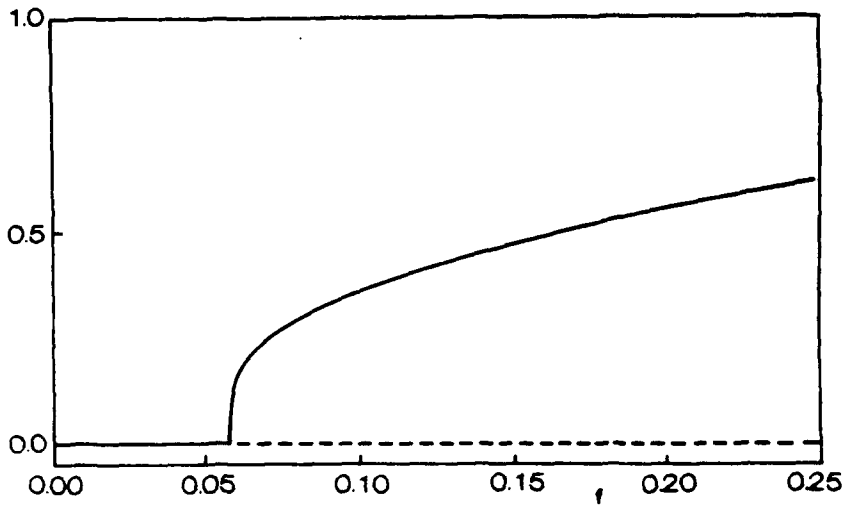


Figure 3. Variation of the steady-state amplitude a* with the amplitude of excitation f in region II of Figure 2: (——) stable, (----) unstable, ϕ = 2.000.

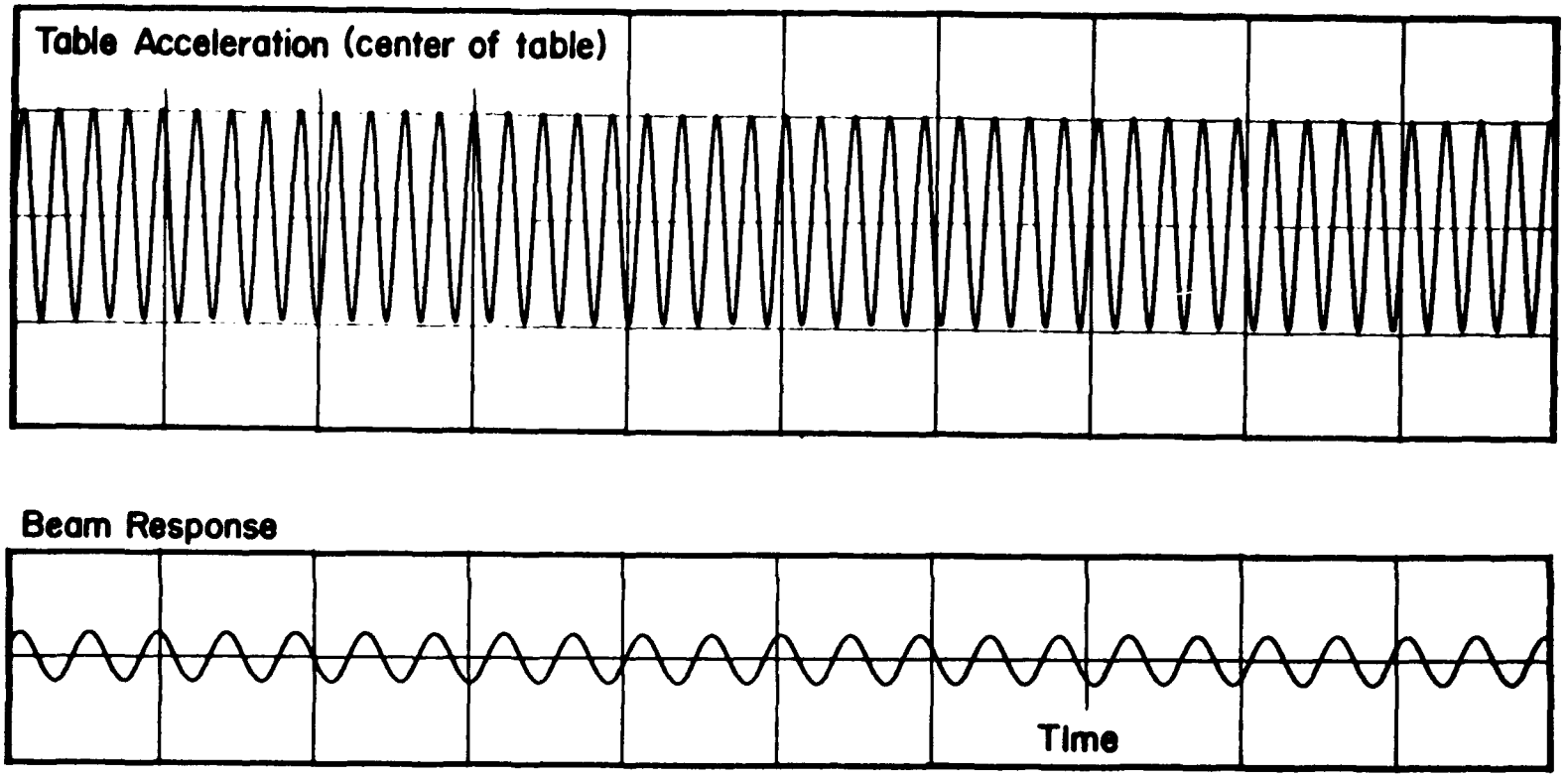


Figure 4. Time history of the shaker table acceleration and the strain gage (beam displacement). Note that the frequency of the response is exactly one-half that of the excitation frequency.

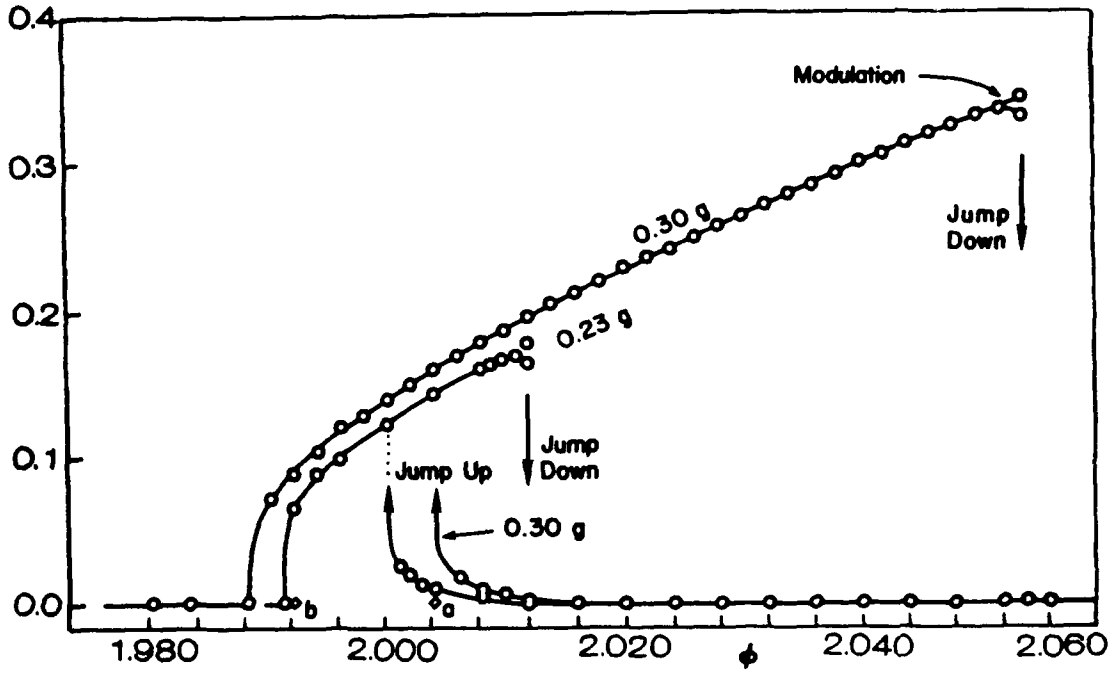


Figure 5. Variation of the steady-state amplitude a* with the frequency of excitation ϕ for beam 1 for two levels of excitation amplitude: 0.230 g's and 0.300 g's. The diamond at "a" denotes the loss of stability of the trivial solution for the 0.300-g acceleration level and the diamond at "b" for the 0.230-g acceleration level. These diamonds represent the extent of penetration into the linear unstable region of the trivial solution.

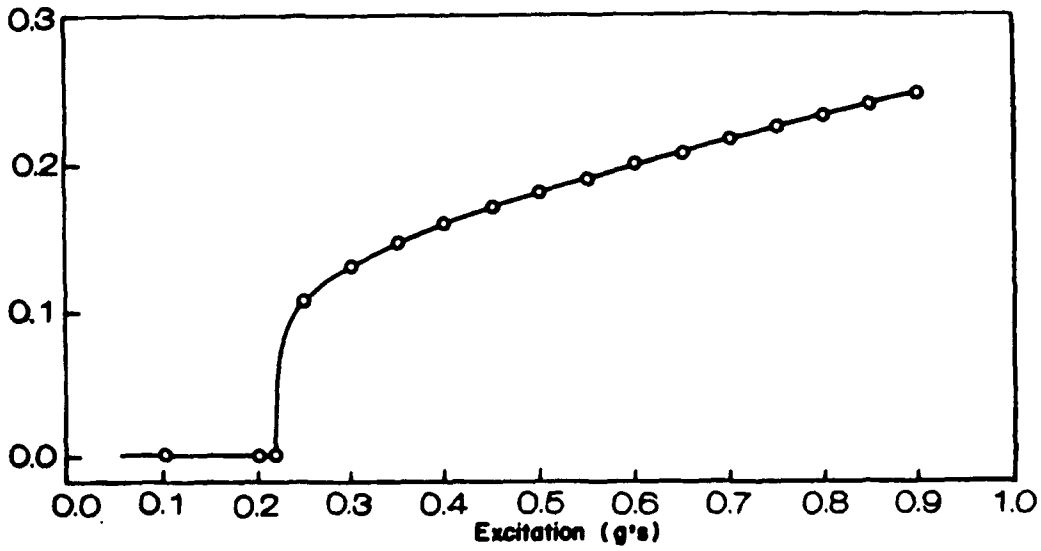


Figure 6. Variation of the amplitude a* with the amplitude of excitation "f" for ϕ = 2.000 for beam 1, shown in Figure 5.

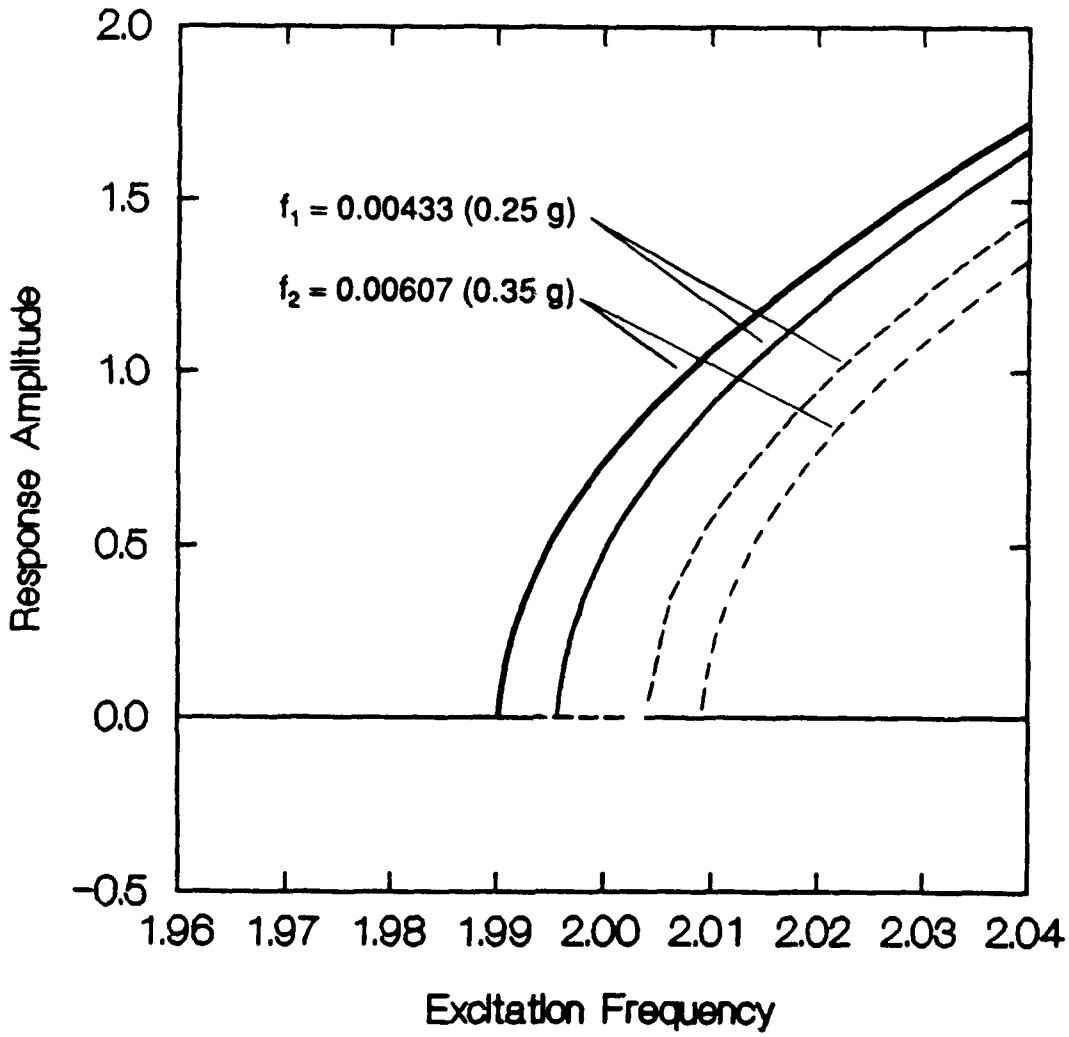


Figure 7. Variation of the steady-state amplitude a* with the nondimensional frequency of excitation of beam 1: (——) stable, (----) unstable, α = 0.3869, κ_1 = 1.8201, κ_2 =0.7379, ζ = 0.001207, v = 0.9585,

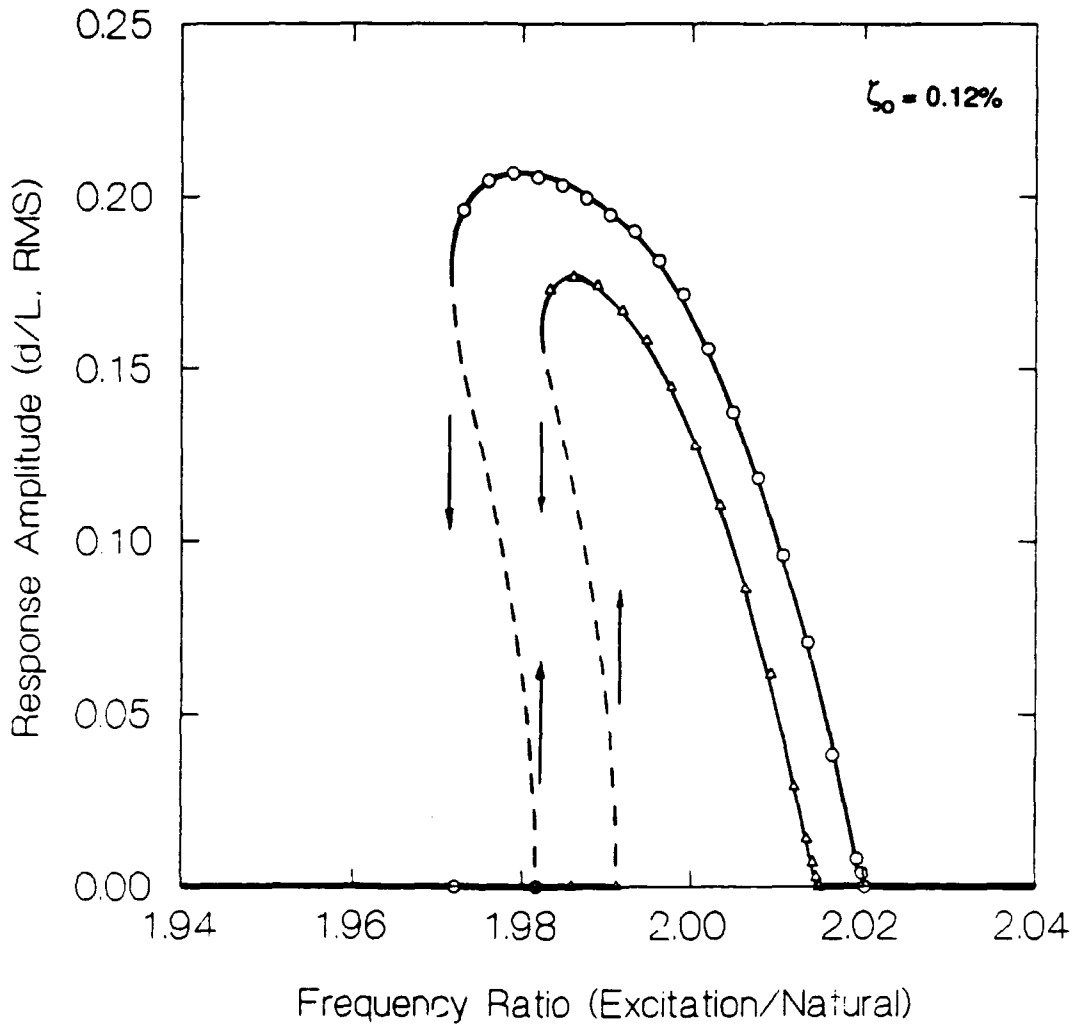


Figure 8. Variation of the steady-state amplitude a* with the frequency of excitation ϕ for the SDOF beam 2 (before damping treatment was applied) for two levels of excitation amplitude: 0.25 g's (v) and 0.35 g's (o) rms. The bending of the curves to the left was neither expected nor predicted--it indicates that the system is softening.

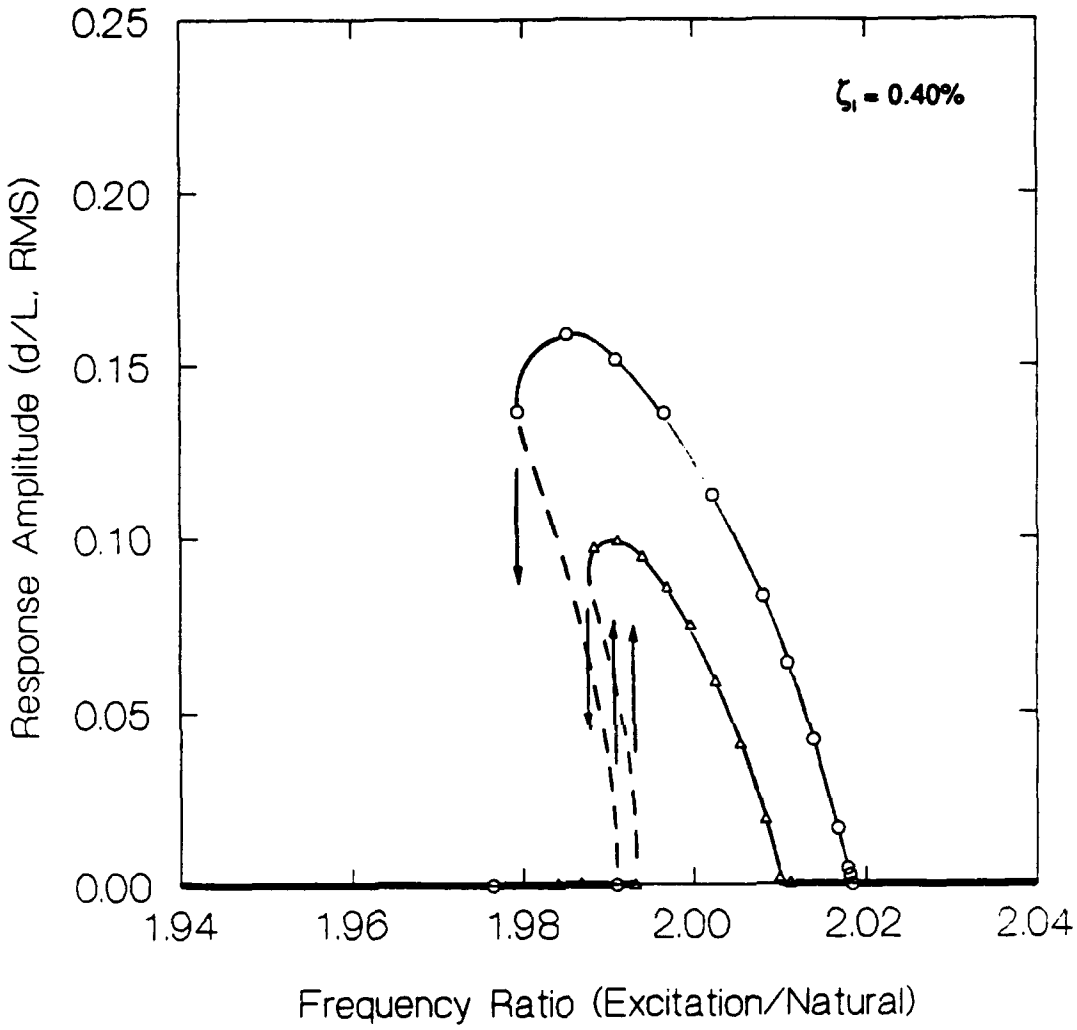


Figure 9. Variation of the steady-state amplitude a* with the frequency of excitation ϕ for two levels of excitation amplitude for beam 2 with one strip of damping treatment applied: 0.25 g's (Δ) and 0.35 g's (o) rms.

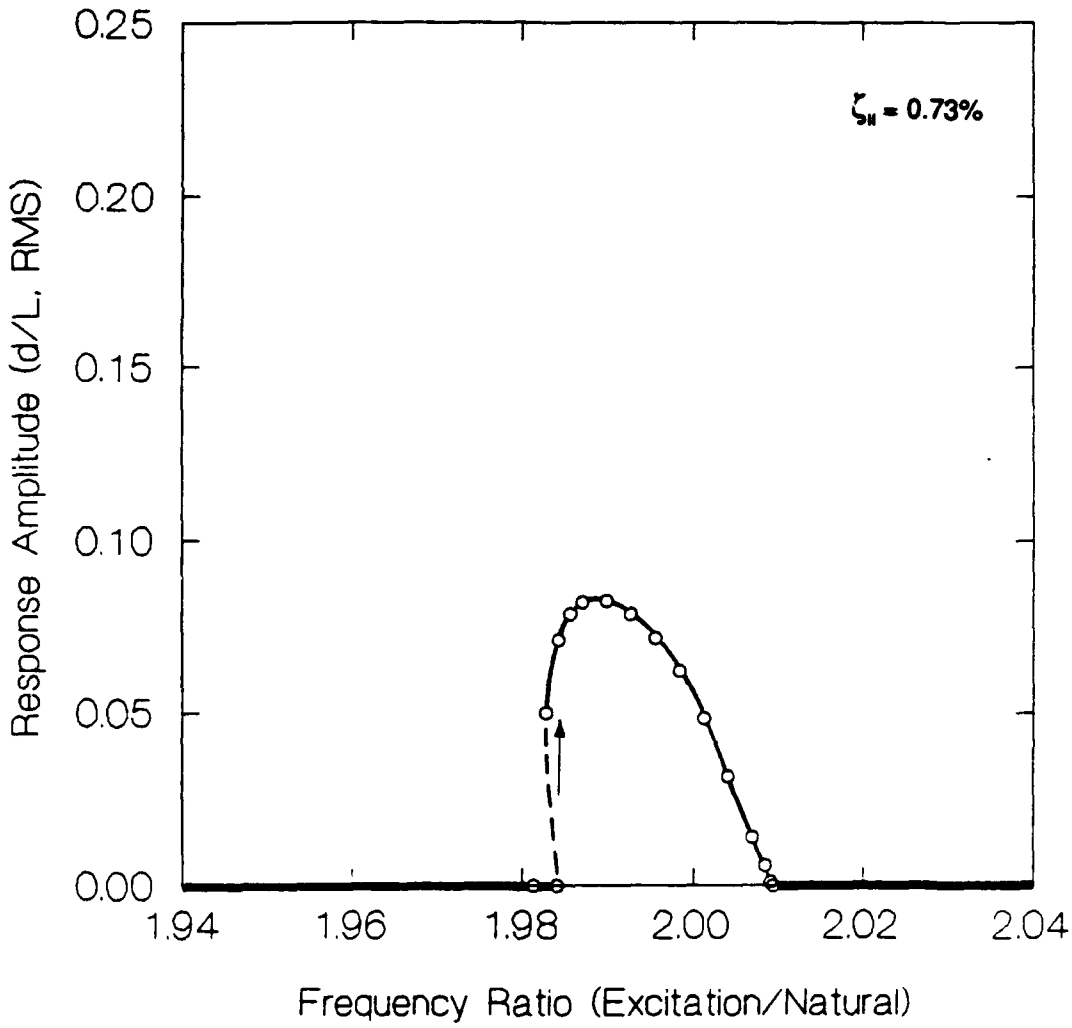


Figure 10. Variation of the steady-state amplitude a* with the frequency of excitation ϕ for two levels of excitation amplitude for beam 2 with two strips of damping treatment applied: 0.25 g's (Δ) and 0.35 g's (o) rms.

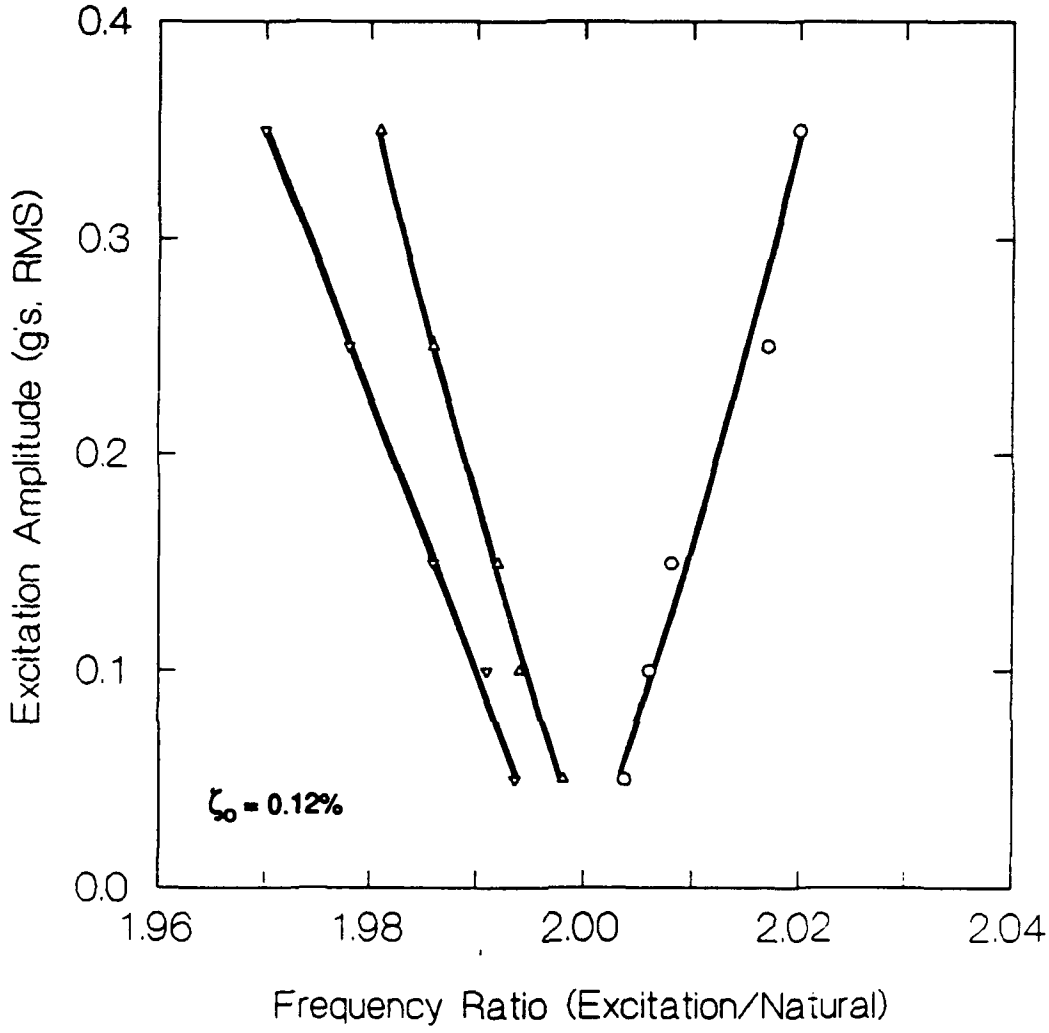


Figure 11. Bifurcation diagram showing the instability regions in the excitation-amplitude vs. excitation frequency domain of the principal parametric resonance for beam 2 before any damping treatment was applied. The curve on the far left represents the boundary caused by the overhang observed in Figure 8. In this case, it represents a sub-critical instability.

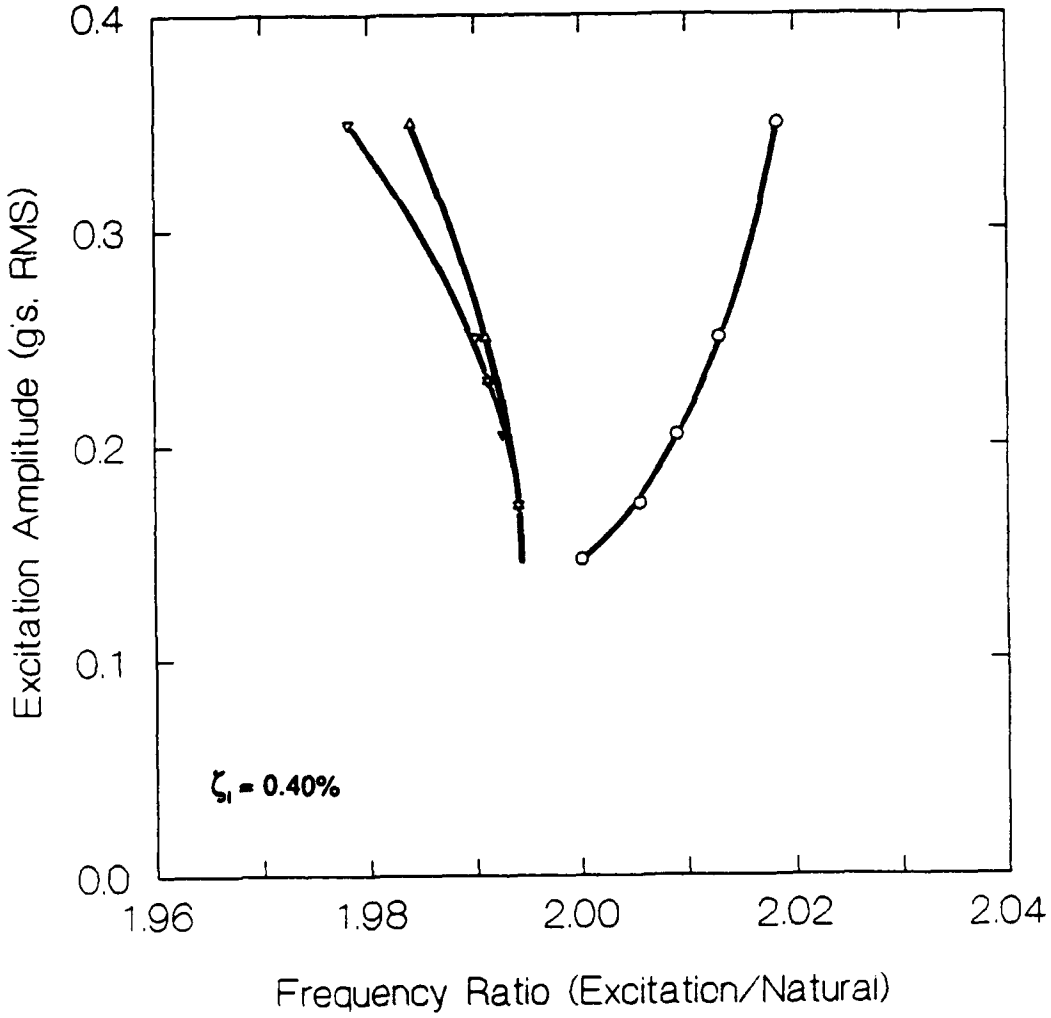


Figure 12. Bifurcation diagram showing the instability regions of the principal parametric resonance for beam 2 with one strip of damping treatment. The curve on the far left represents the boundary caused by the overhang observed in Figure 9.

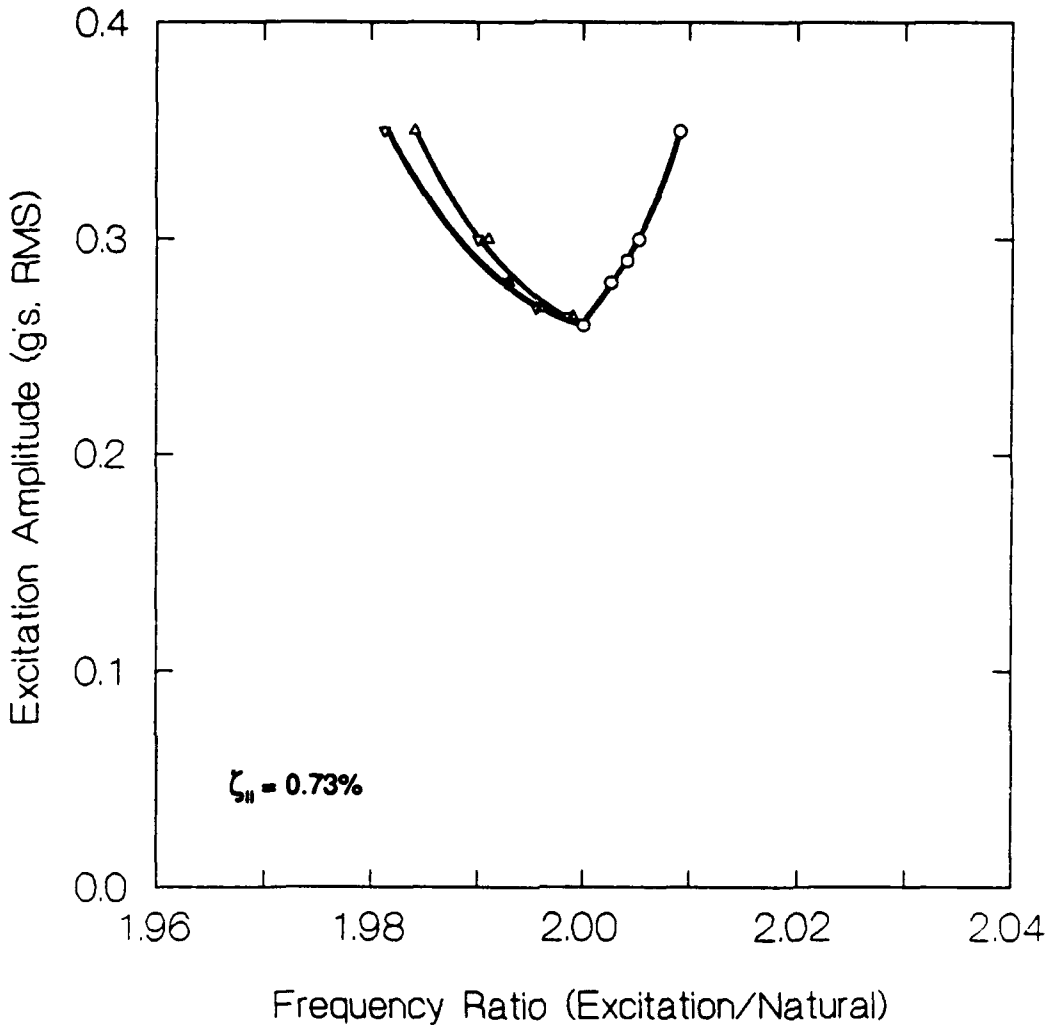


Figure 13. Bifurcation diagram showing the instability regions of the principal parametric resonance for beam 2 with two strips of damping treatment. The curve on the far left represents the boundary caused by the overhang observed in Figure 10.

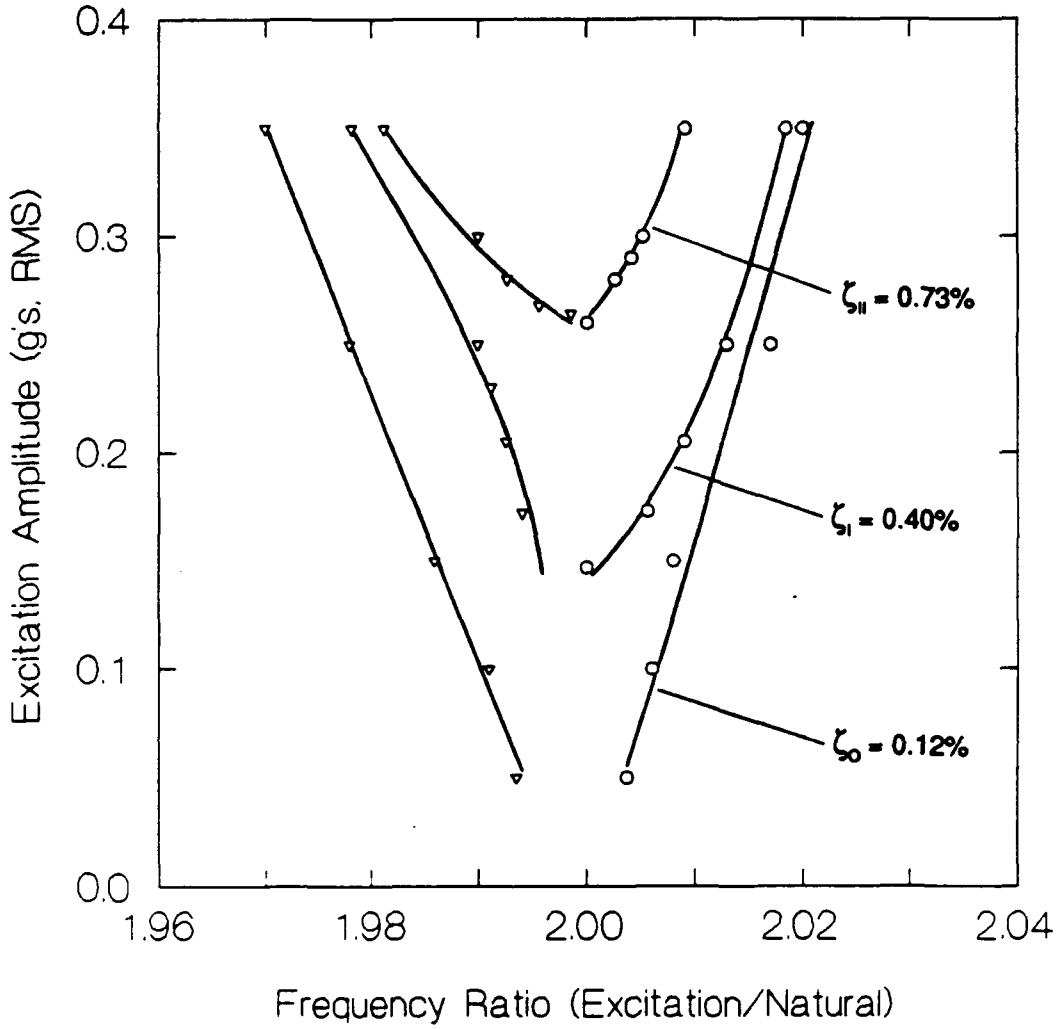


Figure 14. Bifurcation diagram summarizing the region where nontrivial responses exist for three levels of damping. Increased amounts of damping treatment causes the instability region to migrate away from the frequency axis and simultaneously contract.

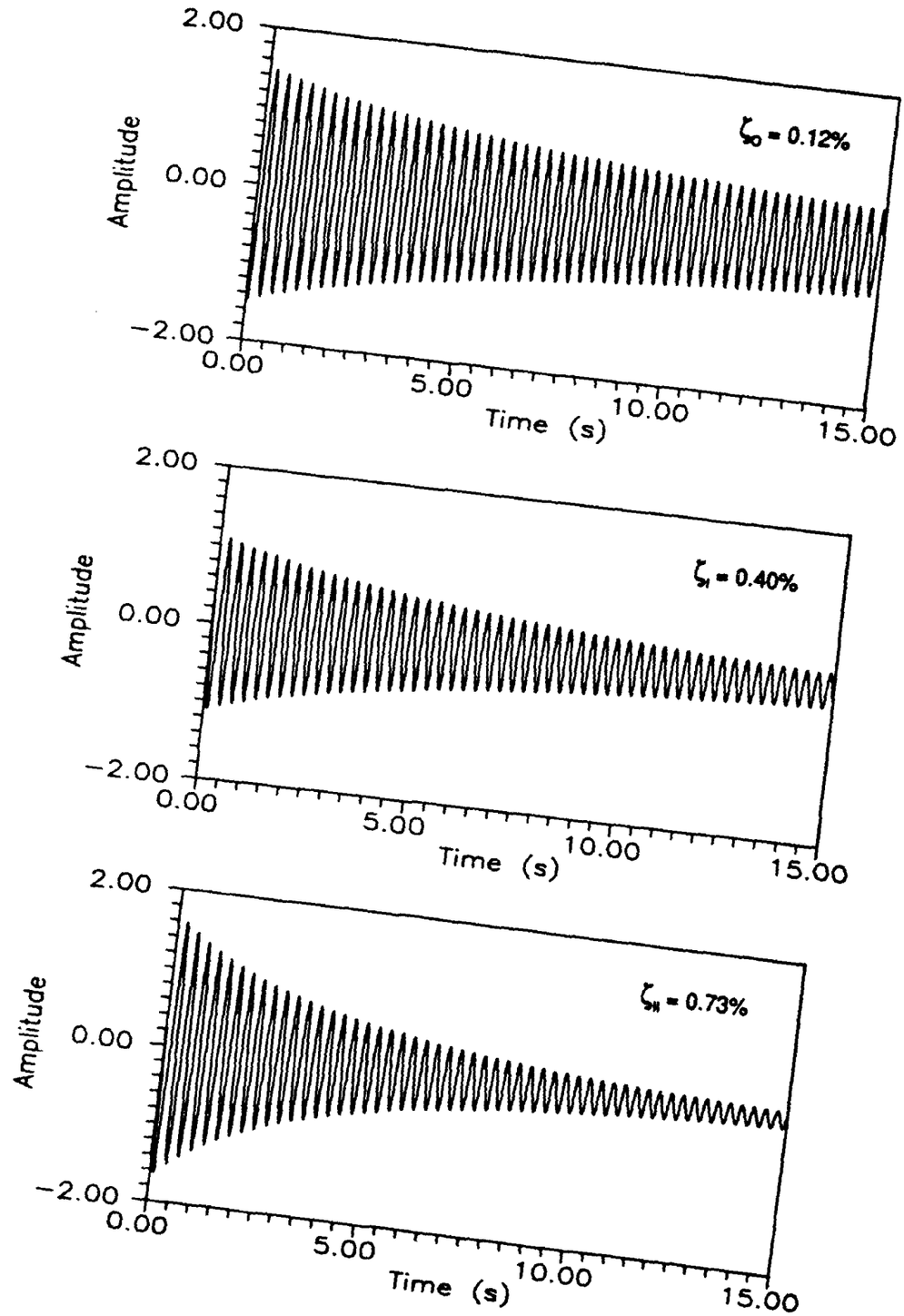


Figure 15. Time histories of the free response from which the natural frequencies and equivalent viscous damping coefficients were computed using the Eigensystem Realization Algorithm (ERA).

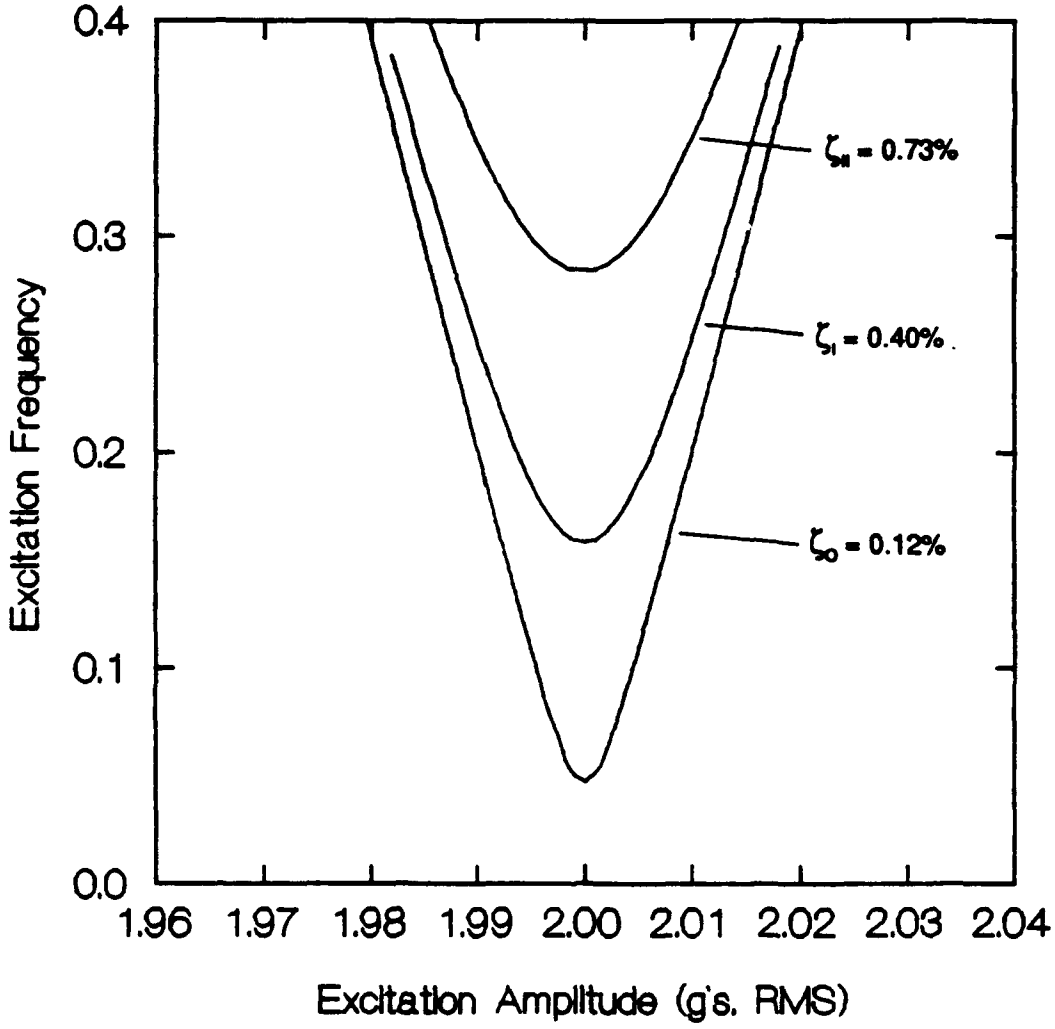


Figure 16. Bifurcation boundaries showing the loss of stability of the trivial solution as predicted by the theory. The damping coefficients were determined from the free decays shown in Figure 15 using ERA.

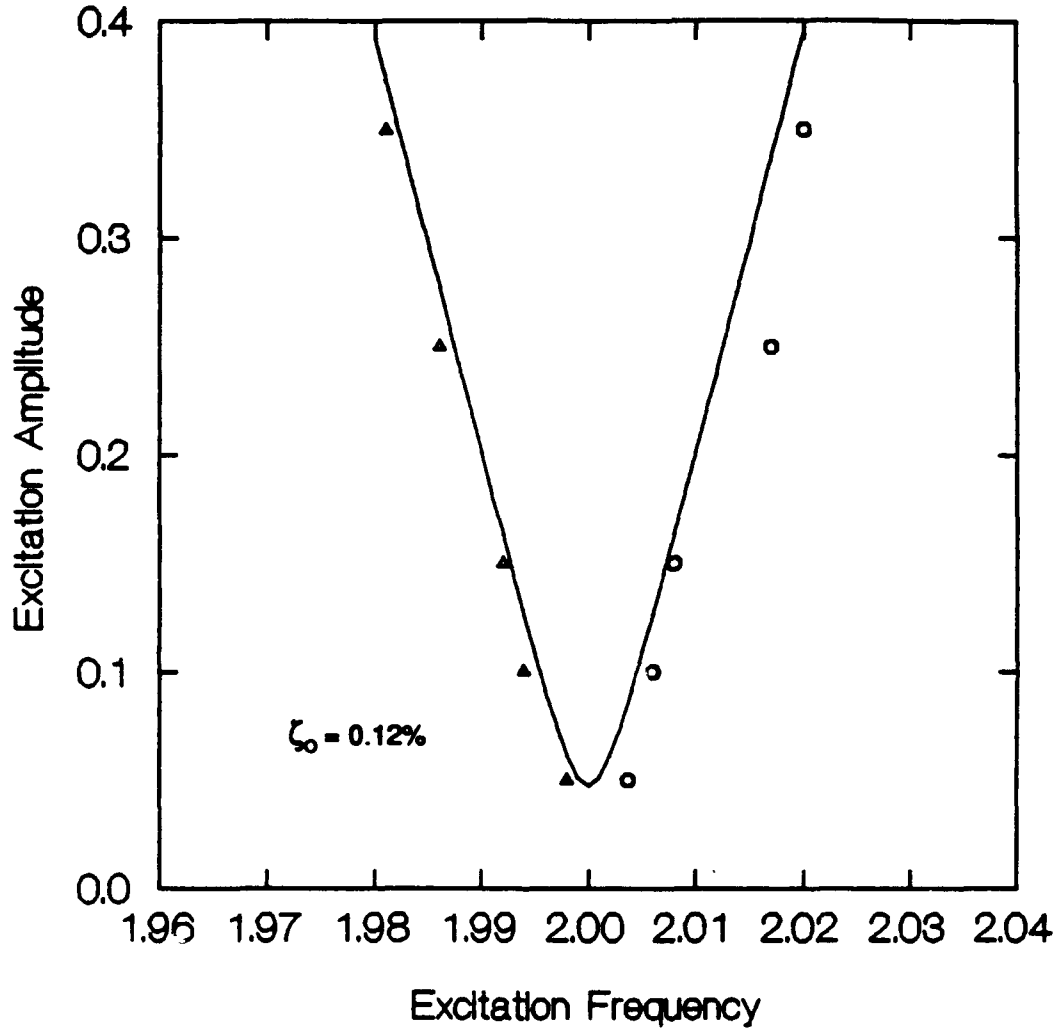


Figure 17. Comparison of the bifurcation boundary indicating the loss of stability of the trivial as predicted by theory and measured during the experiment for the beam with no damping treatment applied: (——) theory, (Δ, o) experiment.

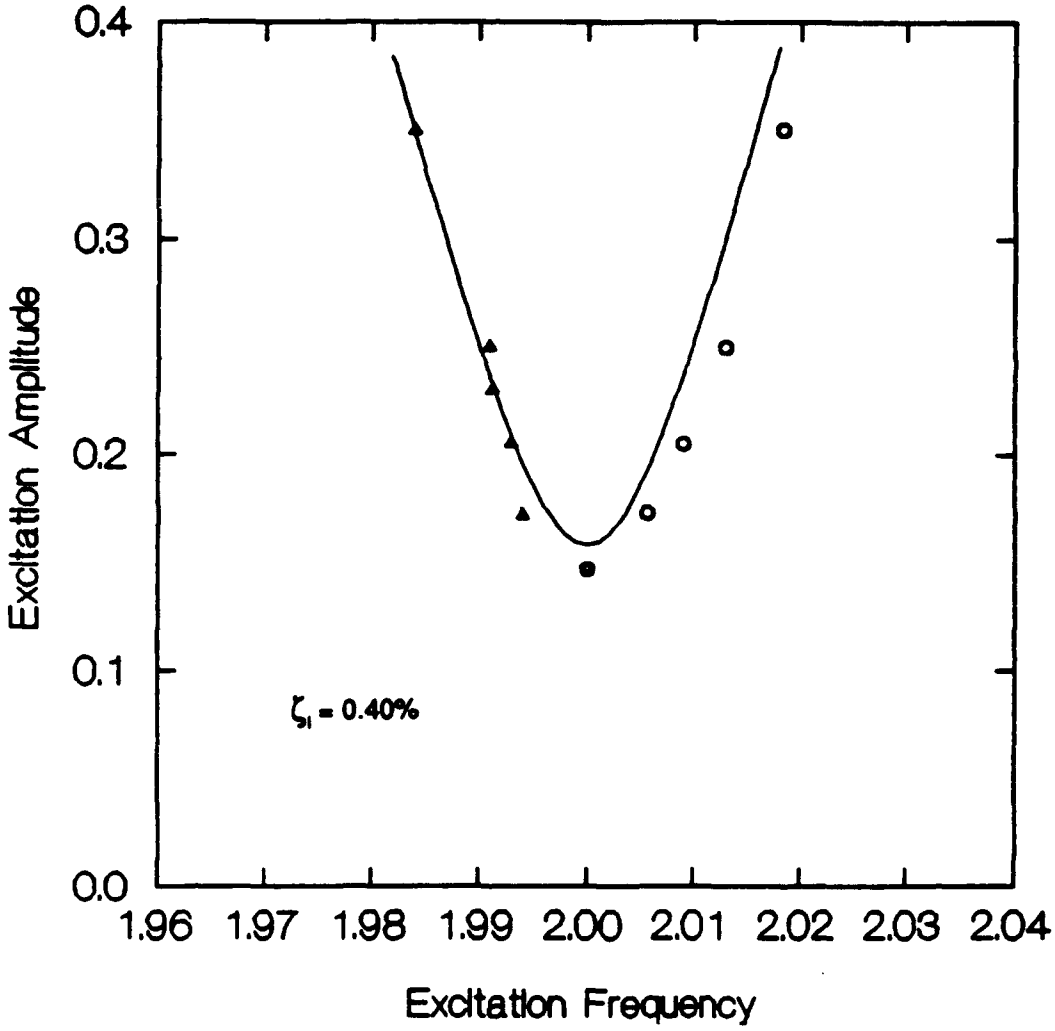


Figure 18. Comparison of the bifurcation boundary indicating the loss of stability of the trivial as predicted by theory and measured during the experiment for the beam with one strip of damping treatment applied: (——) theory, (Δ, o) experiment.

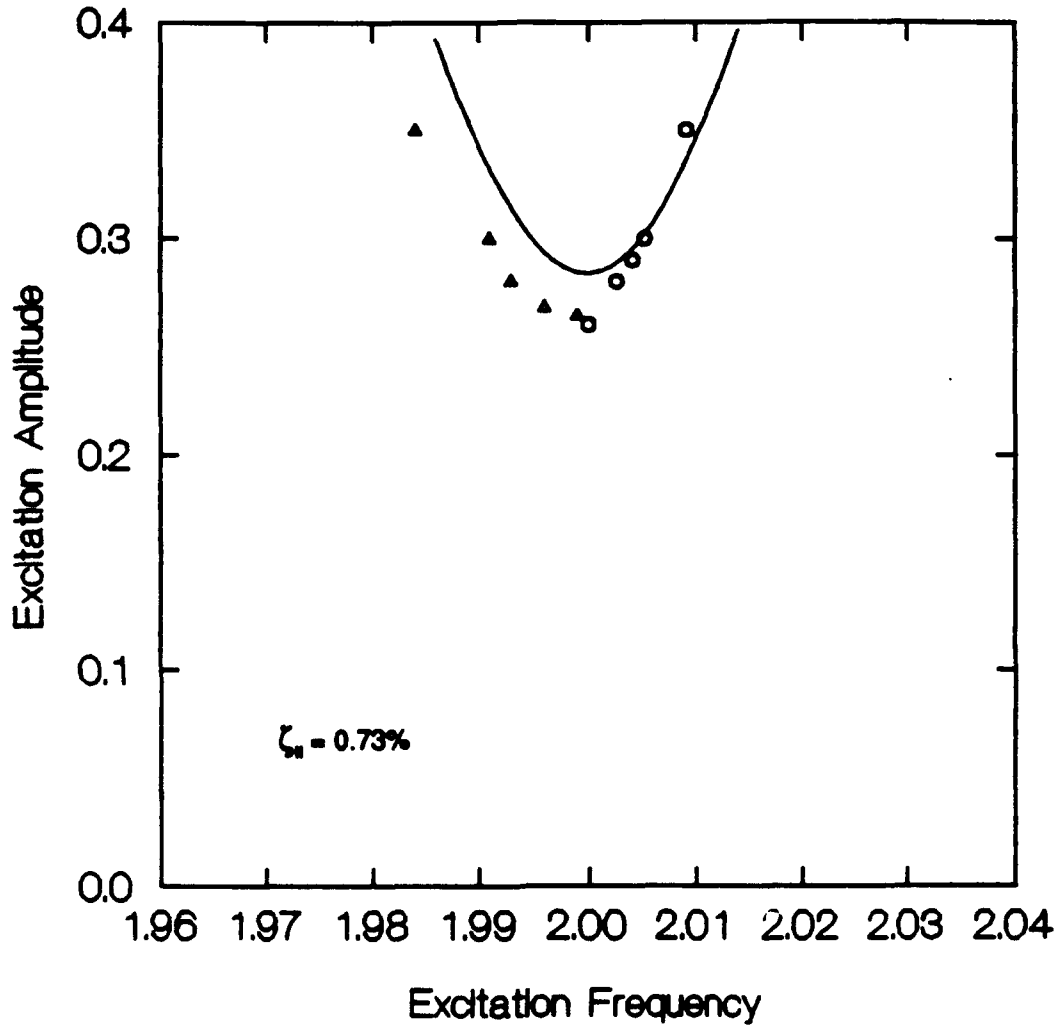


Figure 19. Comparison of the bifurcation boundary indicating the loss of stability of the trivial solution as predicted by theory and measured during the experiment for the beam with two strips of damping treatment applied: (——) theory, (Δ, o) experiment.

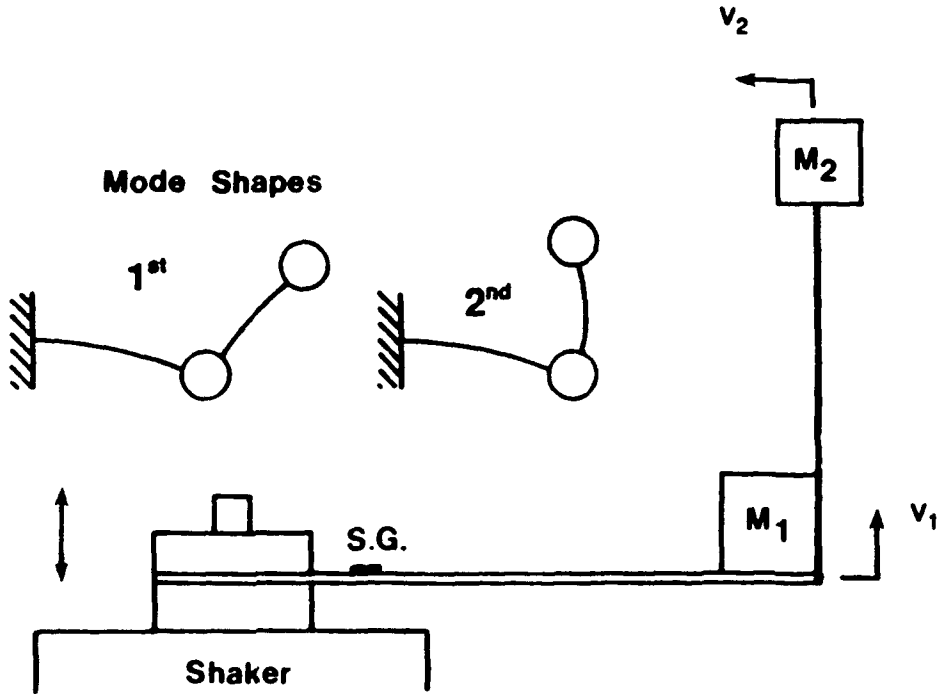


Figure 20. Dual beam-mass structure used for the MDOF experiments. By adjusting the length of the lower beam and the position of the second mass, the first and second resonant frequencies could be adjusted. The higher modes and the out-of-plane modes were not excited during the experiments; hence, a 2DOF mathematical model can be used to adequately describe the observed behavior.

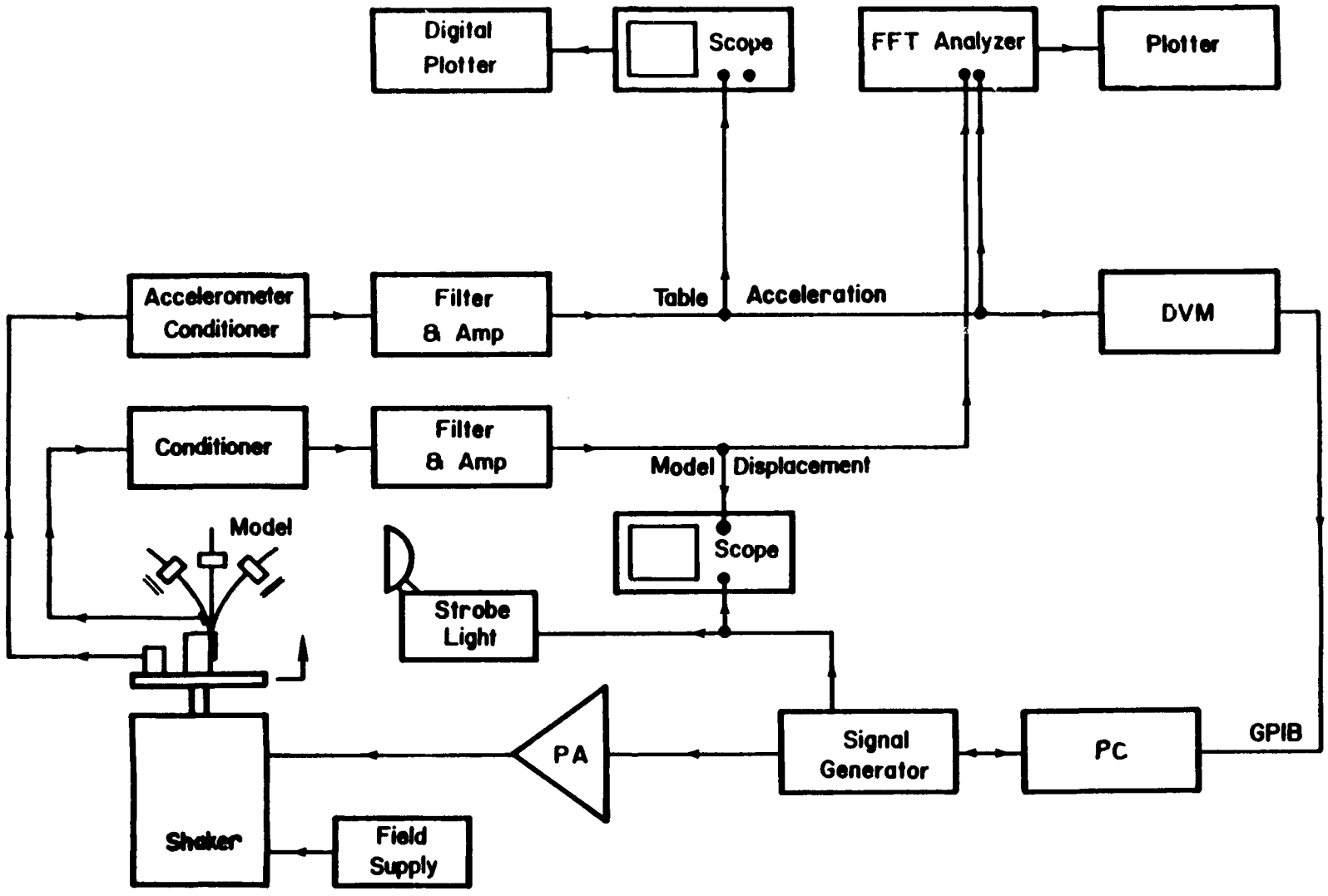


Figure 21. Schematic showing the instrumentation that was used to perform the experiments. Note that a PC was used as a feedback controller to keep the table acceleration amplitude constant during the slow frequency sweeps.

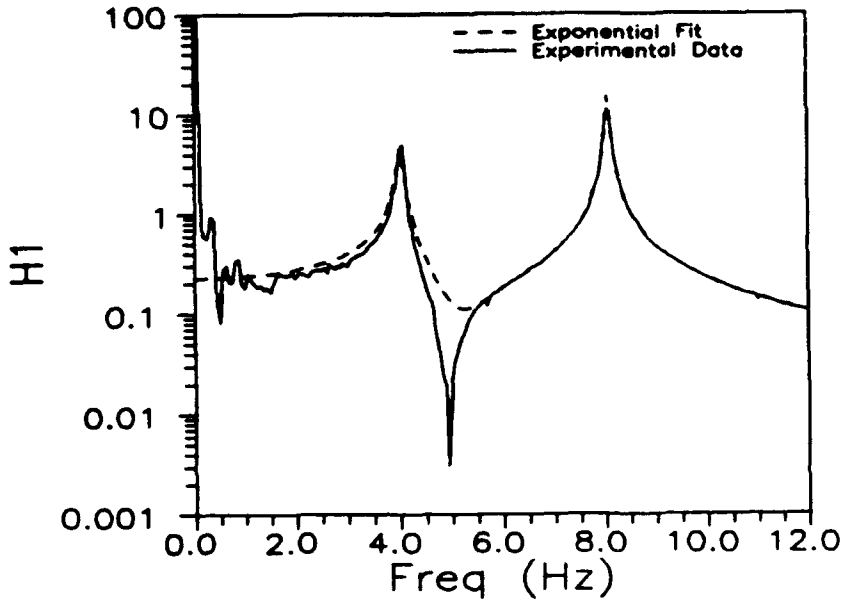


Figure 22. Frequency Response Function and complex exponential curve fit for the tuned structure before damping treatment was applied [O]. The estimated natural frequencies are 4.018 and 8.060 Hz and the damping coefficients are 0.016 and 0.004, respectively. The excitation was low-level random.

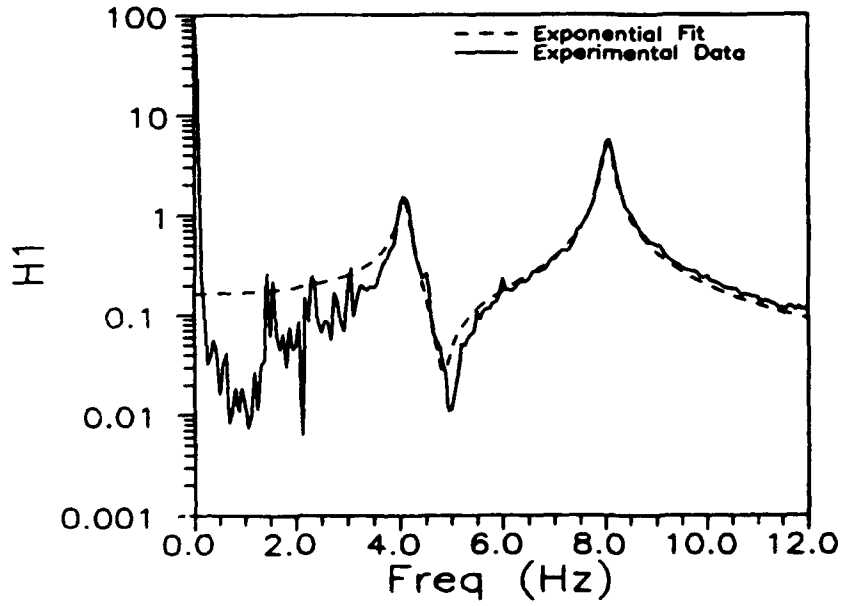


Figure 23. Frequency Response Function and complex exponential curve fit for the tuned structure for the first application of damping treatment [I]. The estimated natural frequencies are 4.091 and 8.050 Hz and the damping coefficients are 0.022 and 0.008, respectively. The excitation was low-level random.

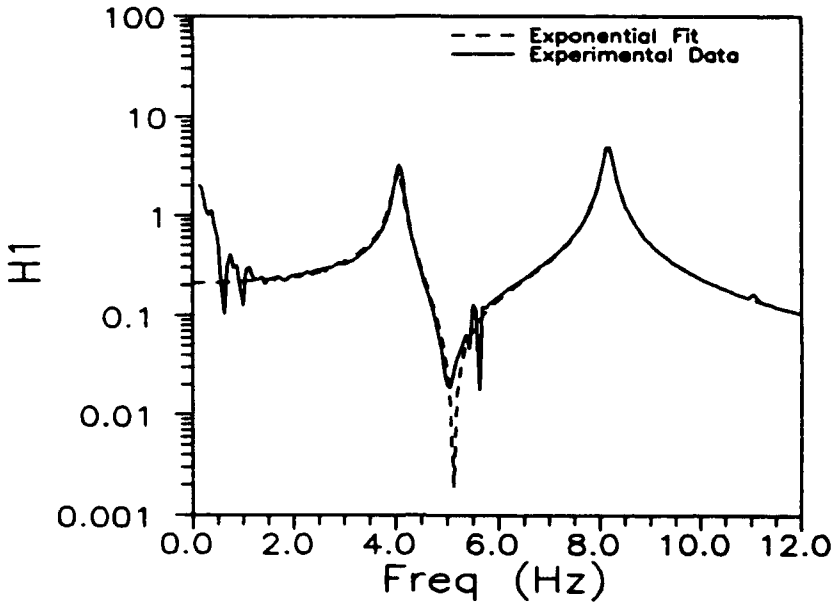


Figure 24. Frequency Response Function and complex exponential curve fit for the tuned structure for the second application of damping treatment [II]. The estimated natural frequencies are 4.062 and 8.161 Hz and the damping coefficients are 0.019 and 0.010, respectively. The excitation was low-level random.

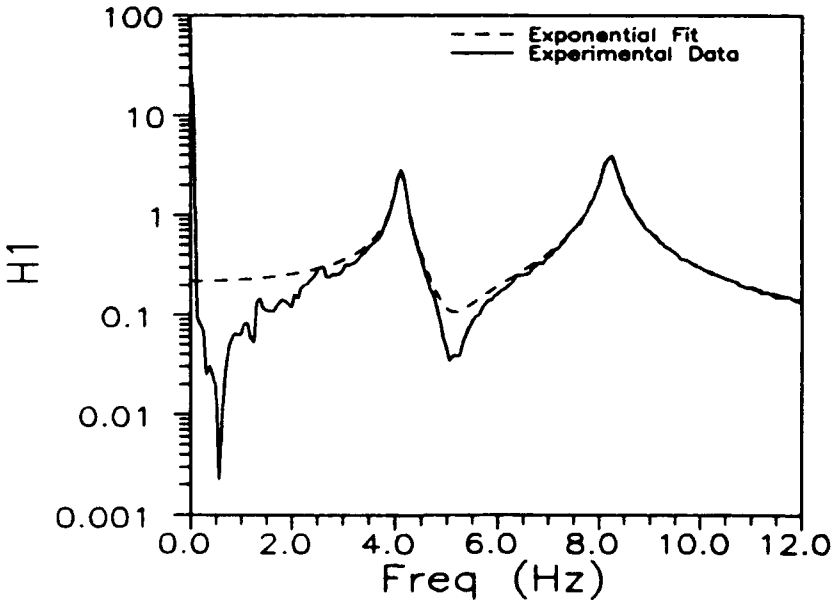


Figure 25. Frequency Response Function and complex exponential curve fit for the tuned structure for the third application of damping treatment [III]. The estimated natural frequencies are 4.131 and 8.220 Hz and the damping coefficients are 0.018 and 0.017, respectively. The excitation was low-level random.

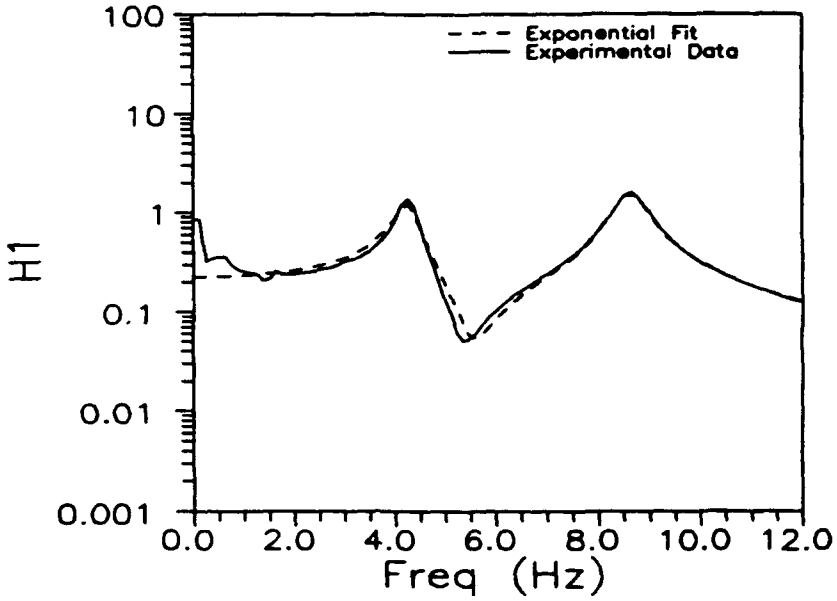


Figure 26. Frequency Response Function and complex exponential curve fit for the tuned structure for the fifth application of damping treatment [V]. The estimated natural frequencies are 4.278 and 8.620 Hz and the damping coefficients are 0.053 and 0.032, respectively. The excitation was low-level random.

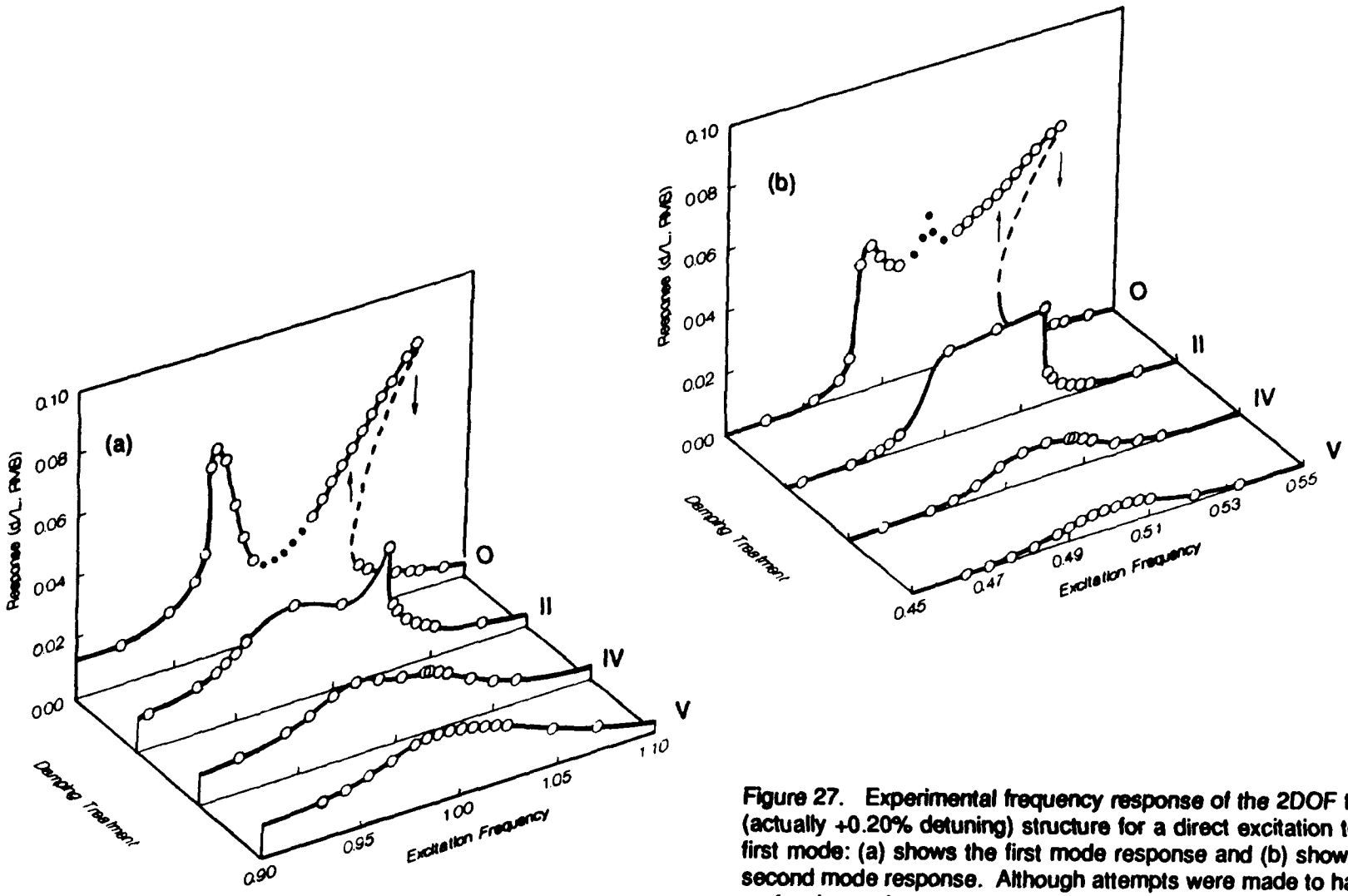


Figure 27. Experimental frequency response of the 2DOF tuned (actually +0.20% detuning) structure for a direct excitation to the first mode: (a) shows the first mode response and (b) shows the second mode response. Although attempts were made to have a perfect internal resonance, there is a small amount of detuning as evidenced from the unsymmetrical nature of the response curves. The arrows indicate jumps.

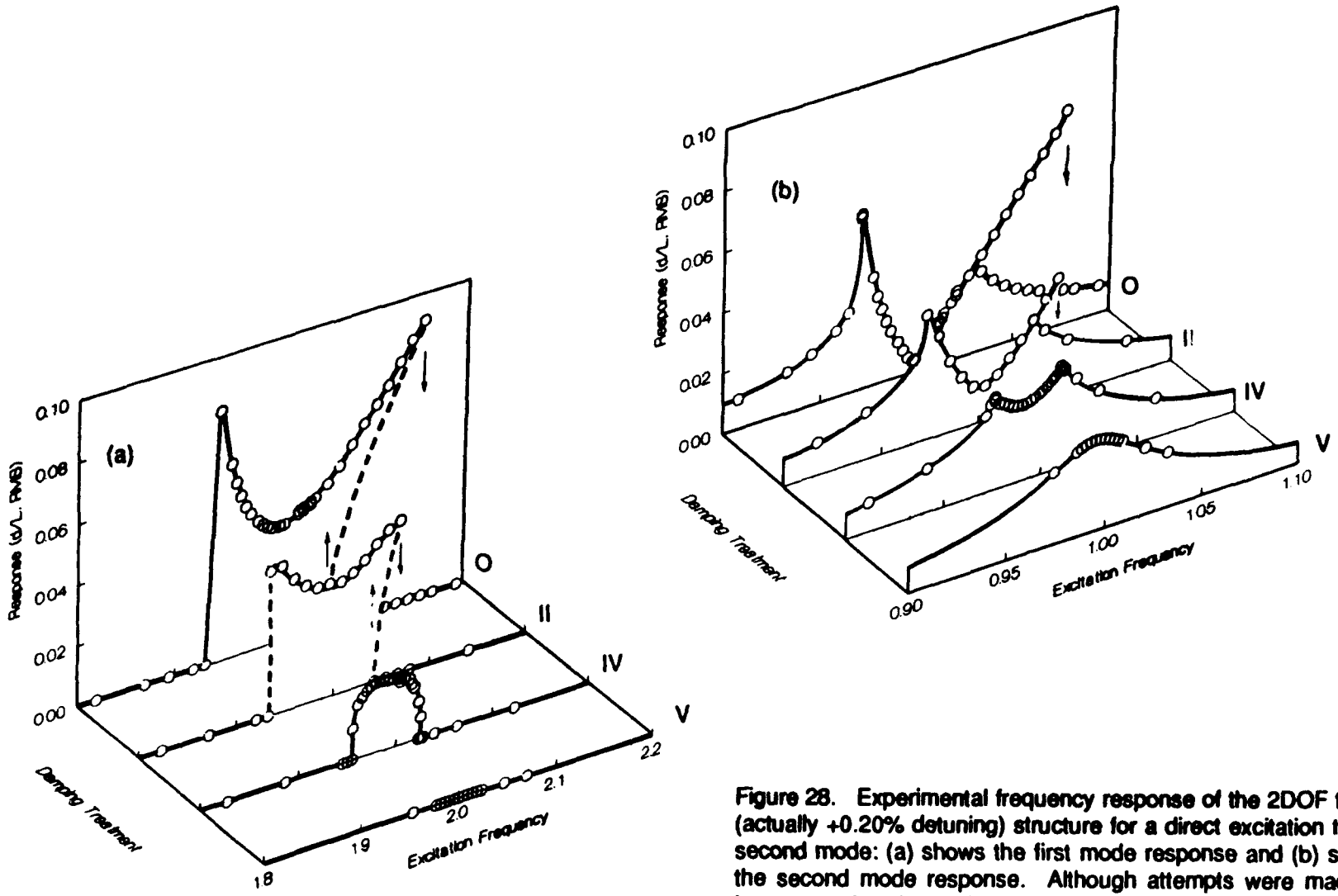


Figure 28. Experimental frequency response of the 2DOF tuned (actually +0.20% detuning) structure for a direct excitation to the second mode: (a) shows the first mode response and (b) shows the second mode response. Although attempts were made to have a perfect internal resonance, there is a small amount of detuning as evidenced from the unsymmetrical nature of the response curves. The arrows indicate jumps.

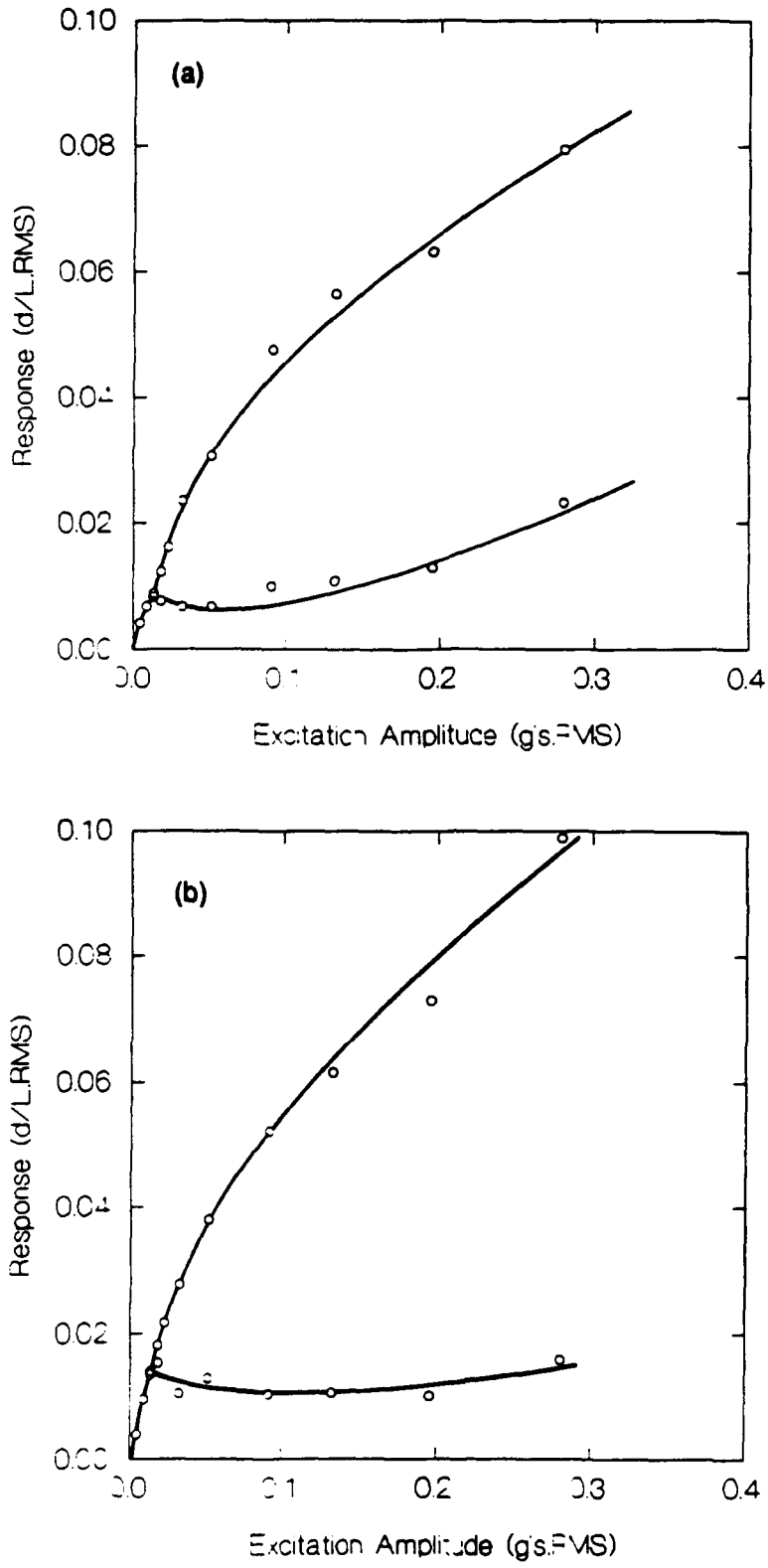
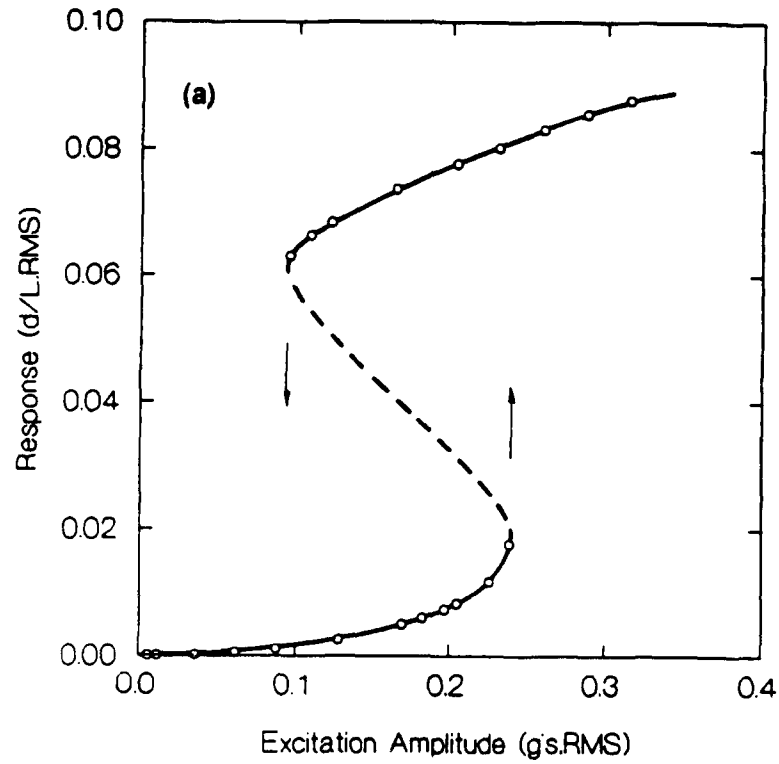


Figure 29. Experimental amplitude response of the 2DOF structure with +0.20% detuning and no damping treatment to a direct excitation to the first mode at a frequency of 1.004: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the modulation region in Figure 27. The curves approximate the bounds on the modulation.



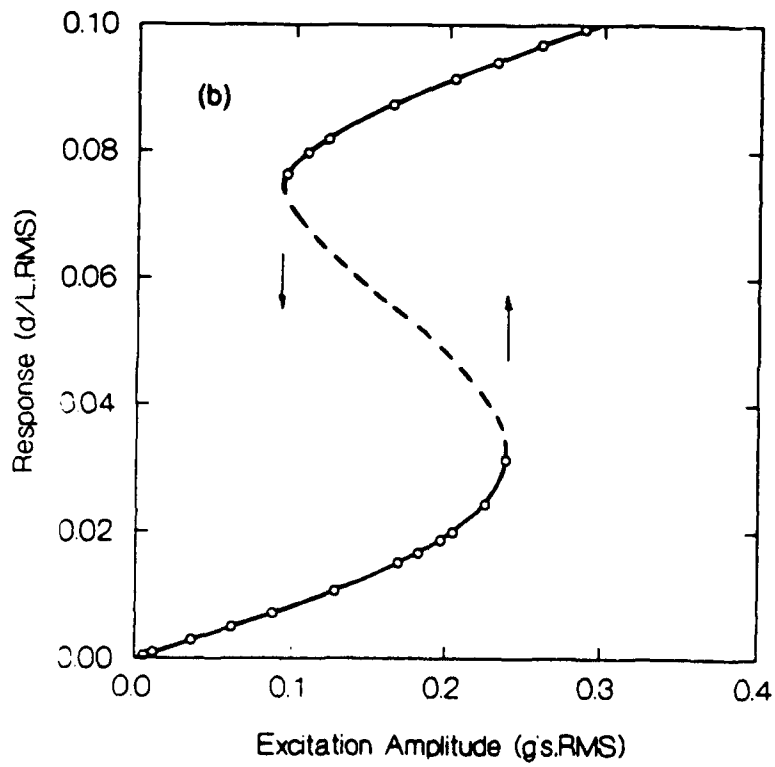
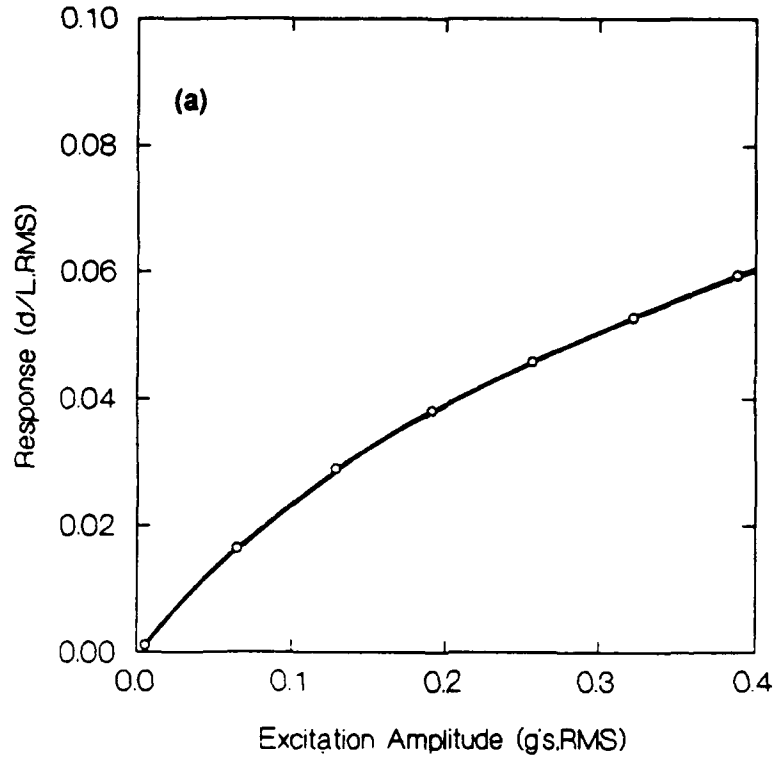


Figure 30. Experimental amplitude response of the 2DOF structure with +0.20% detuning and no damping treatment to a direct excitation to the first mode at a frequency of 1.067: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the double-valued region in Figure 27 (to the right of the modulation region).



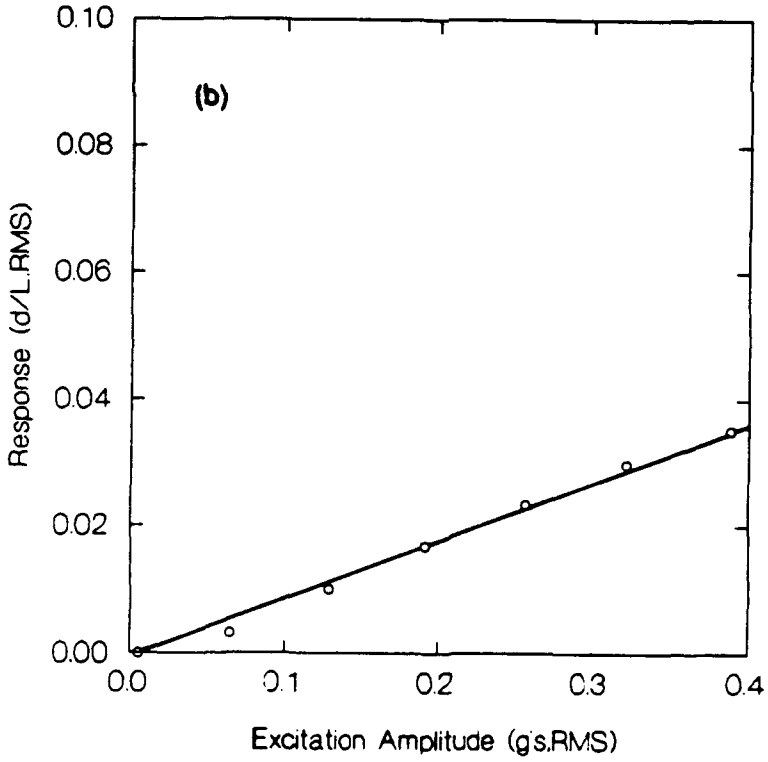


Figure 31. Experimental amplitude response of the 2DOF structure with +0.29% detuning and maximum damping treatment (case [V])to a direct excitation to the first mode at a frequency of 0.979: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the resonance region in Figure 27 [V]. These curves should be compared to Figure 27 to see the effect of the damping treatment--it completely suppresses the modulation and attenuates the amplitude of the response.

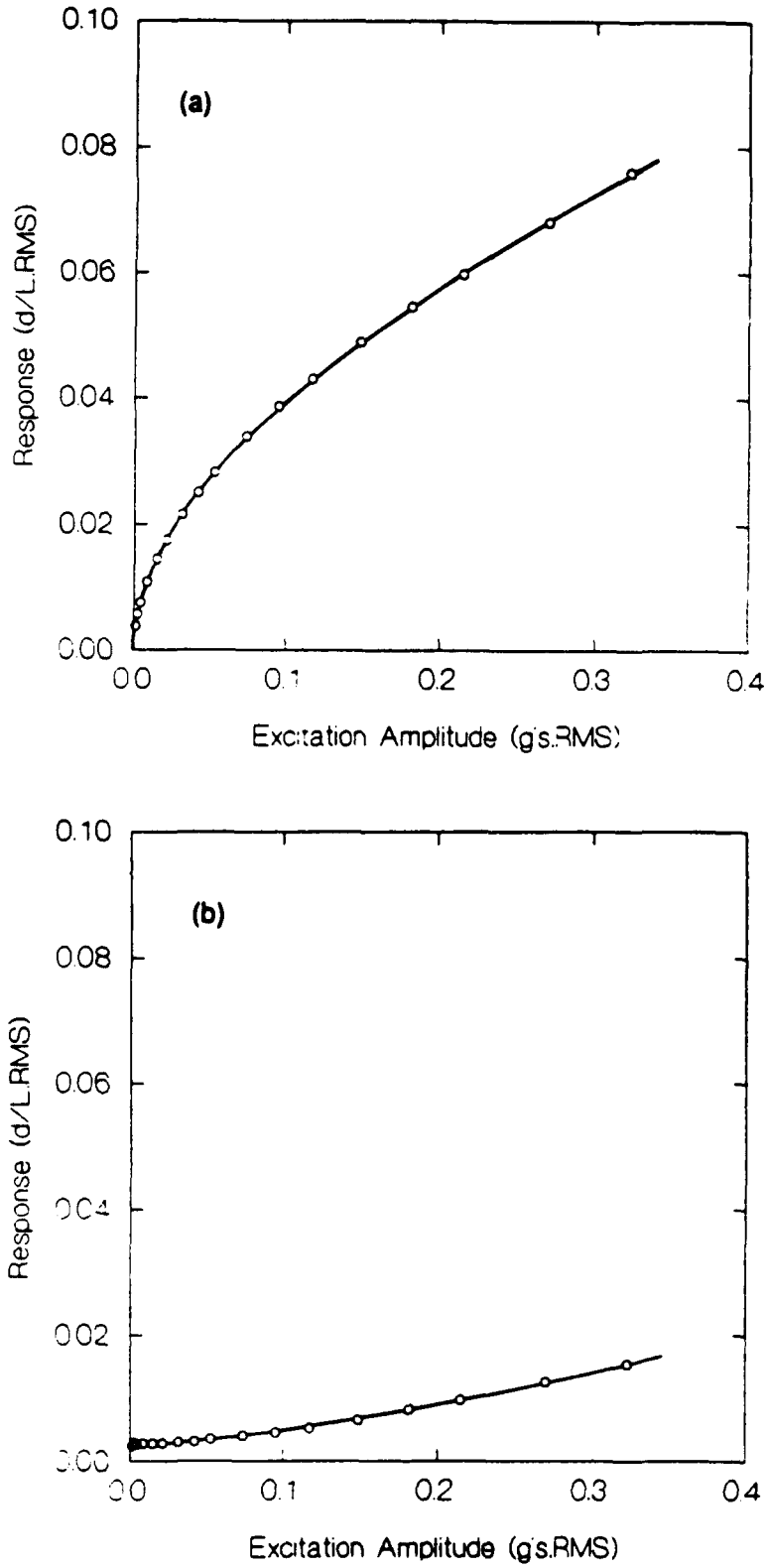


Figure 32. Experimental amplitude response of the 2DOF structure with +0.20% detuning and no damping treatment to a direct excitation to the second mode at a frequency of 0.998: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the single-valued resonance region in Figure 28.

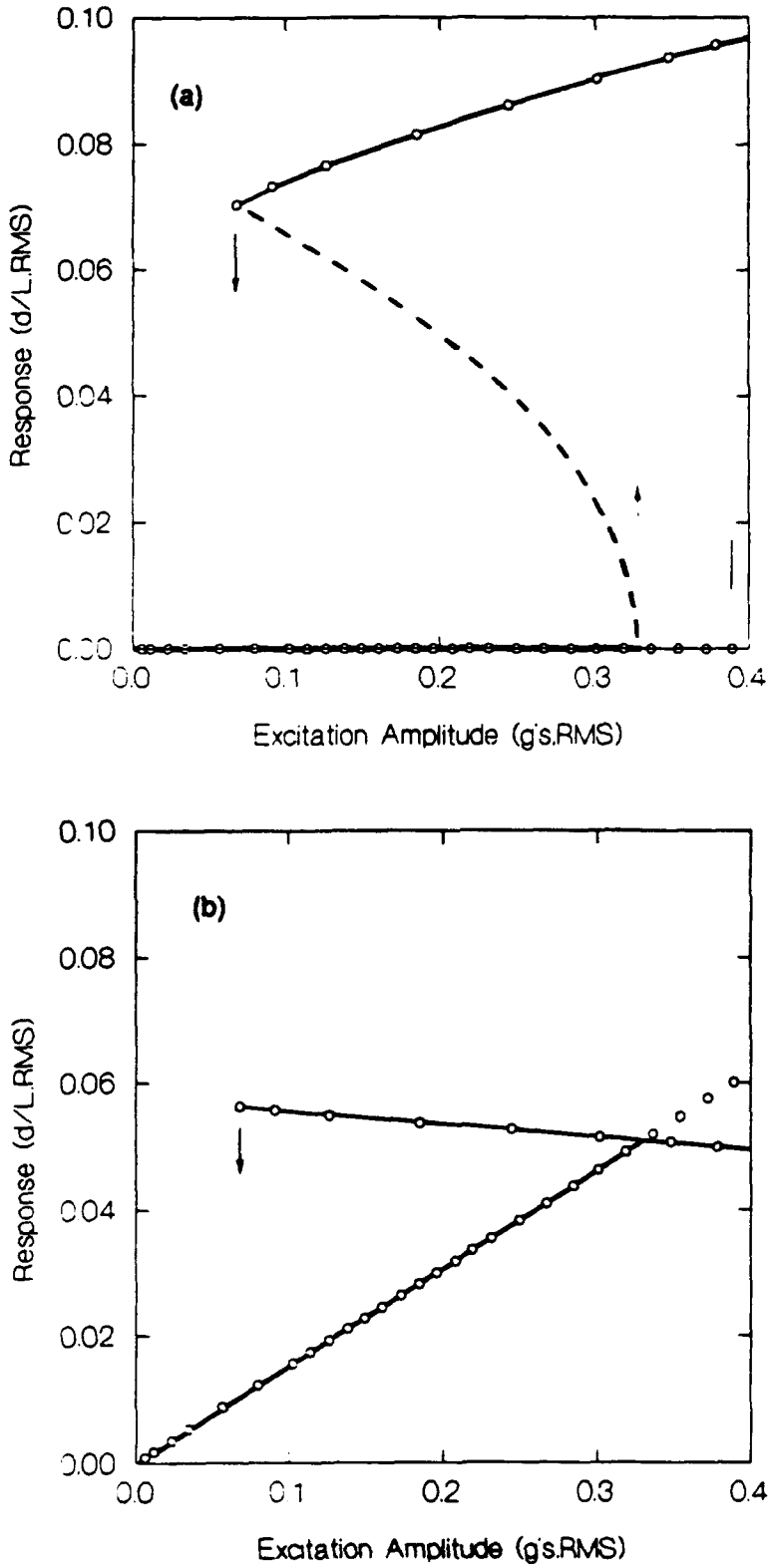


Figure 33. Experimental amplitude response of the 2DOF structure with +0.20% detuning and no damping treatment to a direct excitation to the second mode at a frequency of 1.057: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the overhang region in Figure 28.

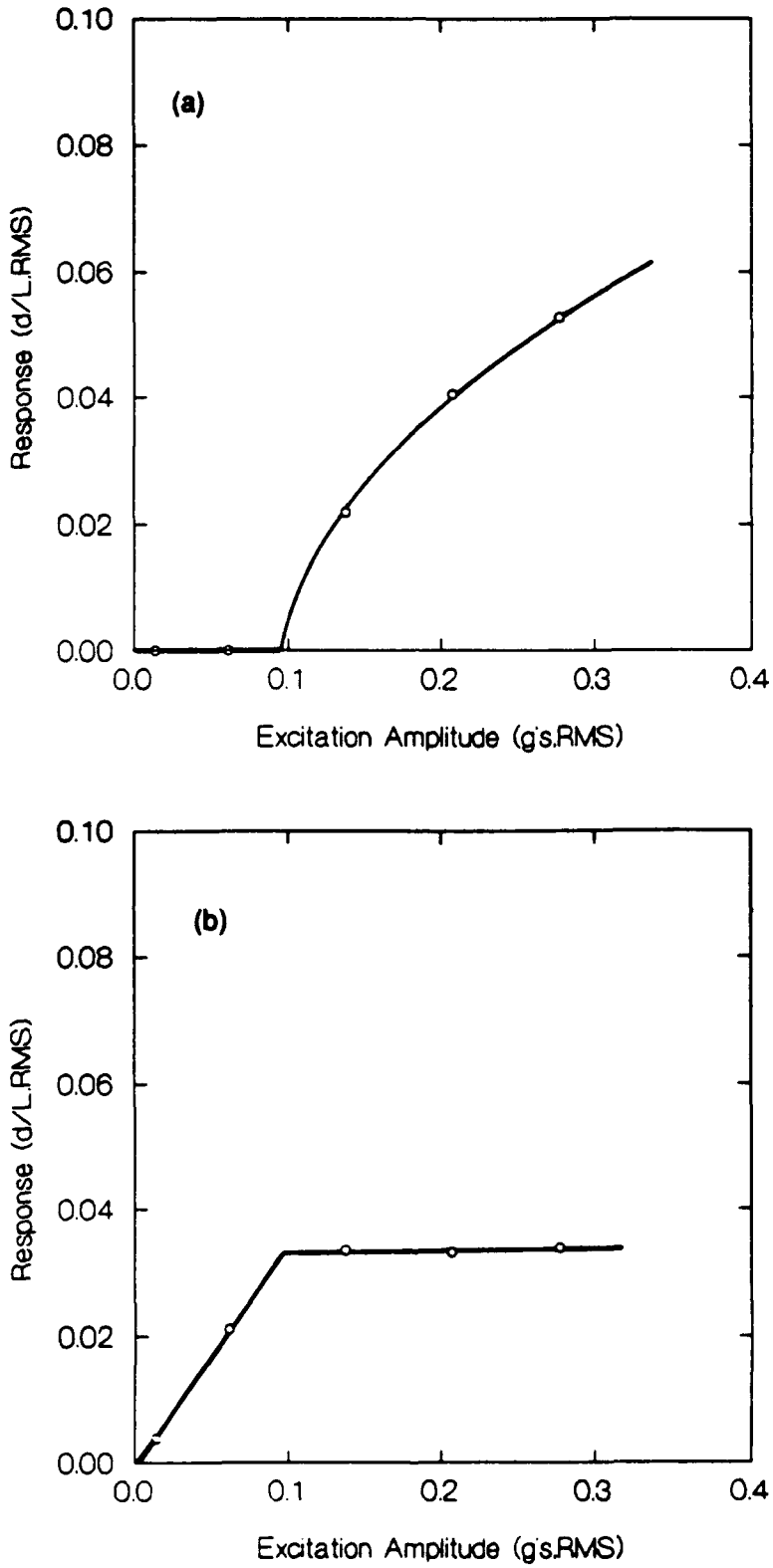


Figure 34. Experimental amplitude response of the 2DOF structure with +0.29% detuning and maximum damping treatment (case [V]) to a direct excitation to the second mode at a frequency of 0.992: (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the peak in Figure 28 [V].

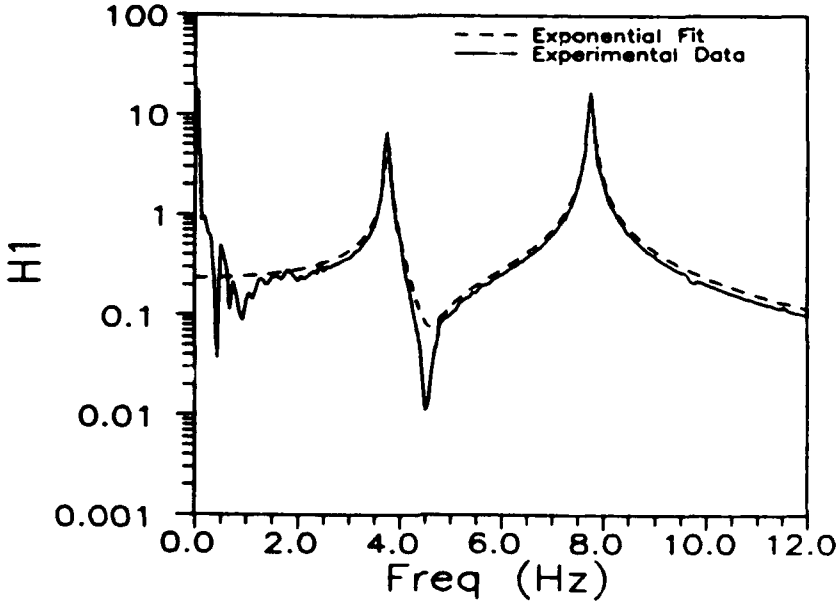


Figure 35. Frequency Response Function and complex exponential curve fit for the slightly detuned structure before damping treatment was applied (case O). The estimated natural frequencies are 3.7599 and 7.7576 Hz and the damping coefficients are 0.008 and 0.004, respectively. The excitation was low-level random.

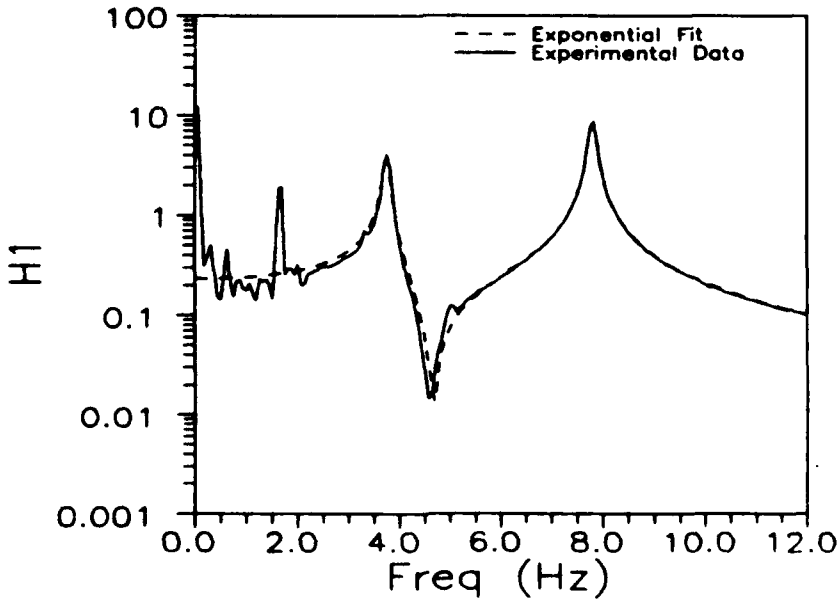


Figure 36. Frequency Response Function and complex exponential curve fit for the slightly detuned structure for damping treatment I. The estimated natural frequencies are 3.7484 and 7.8015 and the damping coefficients are 0.012 and 0.007, respectively. The excitation was low-level random.

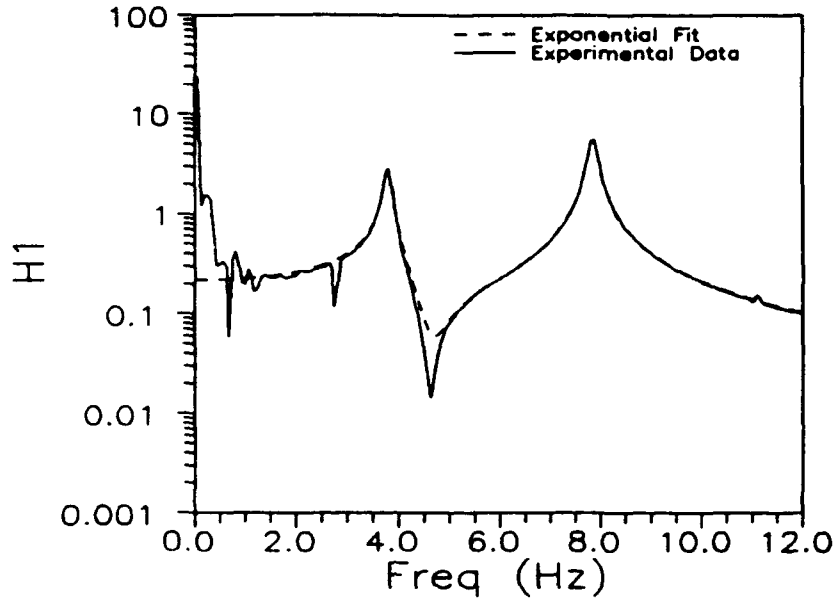


Figure 37. Frequency Response Function and complex exponential curve fit for the slightly detuned structure for damping treatment II. The estimated natural frequencies are 3.7954 and 7.8460 Hz and the damping coefficients are 0.016 and 0.010, respectively. The excitation was low-level random.

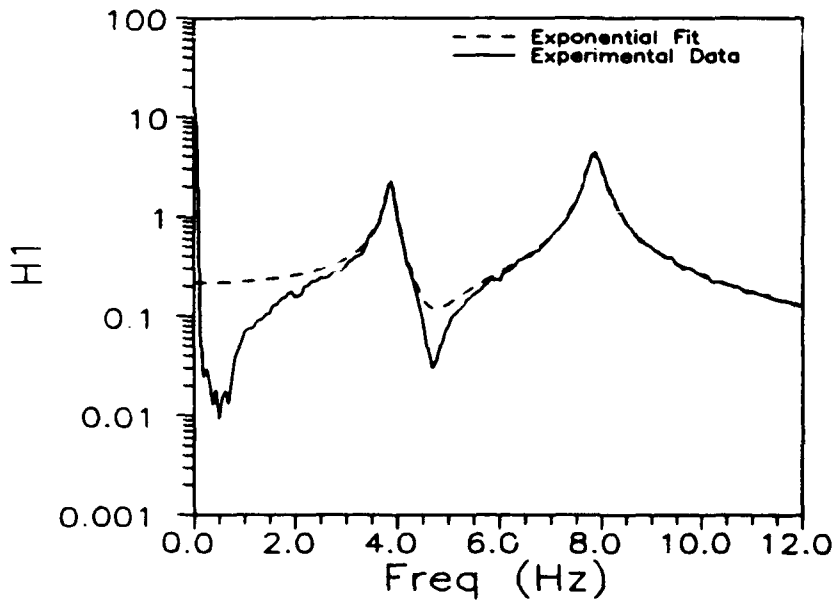


Figure 38. Frequency Response Function and complex exponential curve fit for the slightly detuned structure for damping treatment III. The estimated natural frequencies are 3.8633 and 7.8607 Hz and the damping coefficients are 0.021 and 0.016, respectively. The excitation was low-level random.

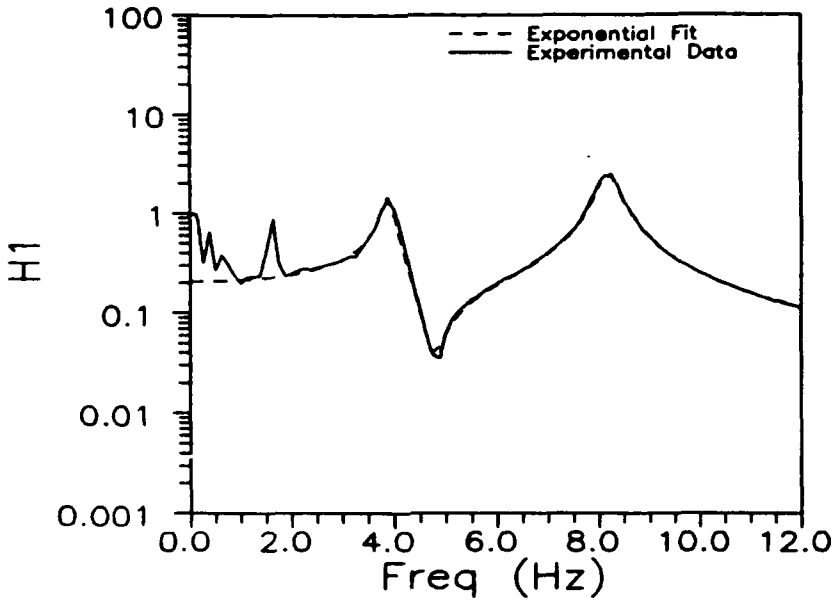


Figure 39. Frequency Response Function and complex exponential curve fit for the slightly detuned structure for damping treatment IV. The estimated natural frequencies are 3.9063 and 8.1856 Hz and the damping coefficients are 0.030 and 0.023, respectively. The excitation was low-level random.

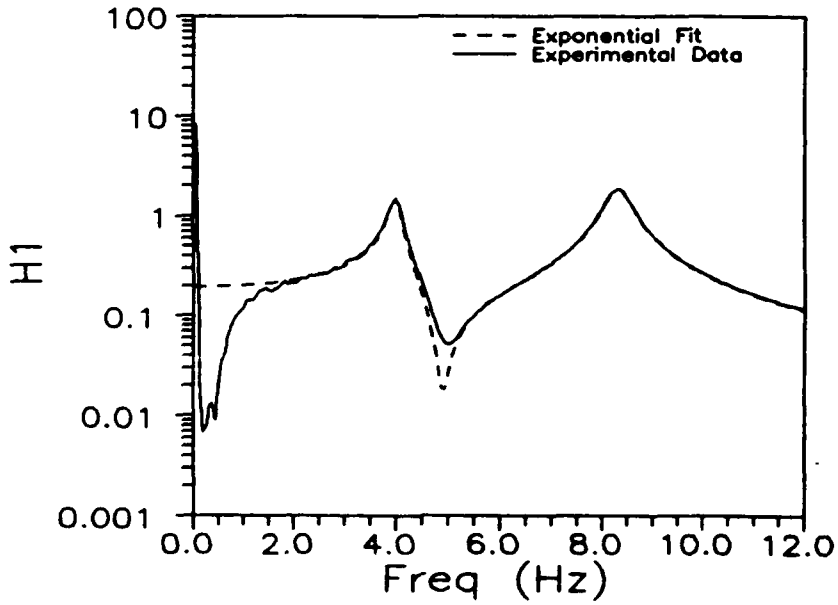


Figure 40. Frequency Response Function and complex exponential curve fit for the slightly detuned structure for damping treatment V. The estimated natural frequencies are 4.0054 and 8.3241 Hz and the damping coefficients are 0.029 and 0.030, respectively. The excitation was low-level random.

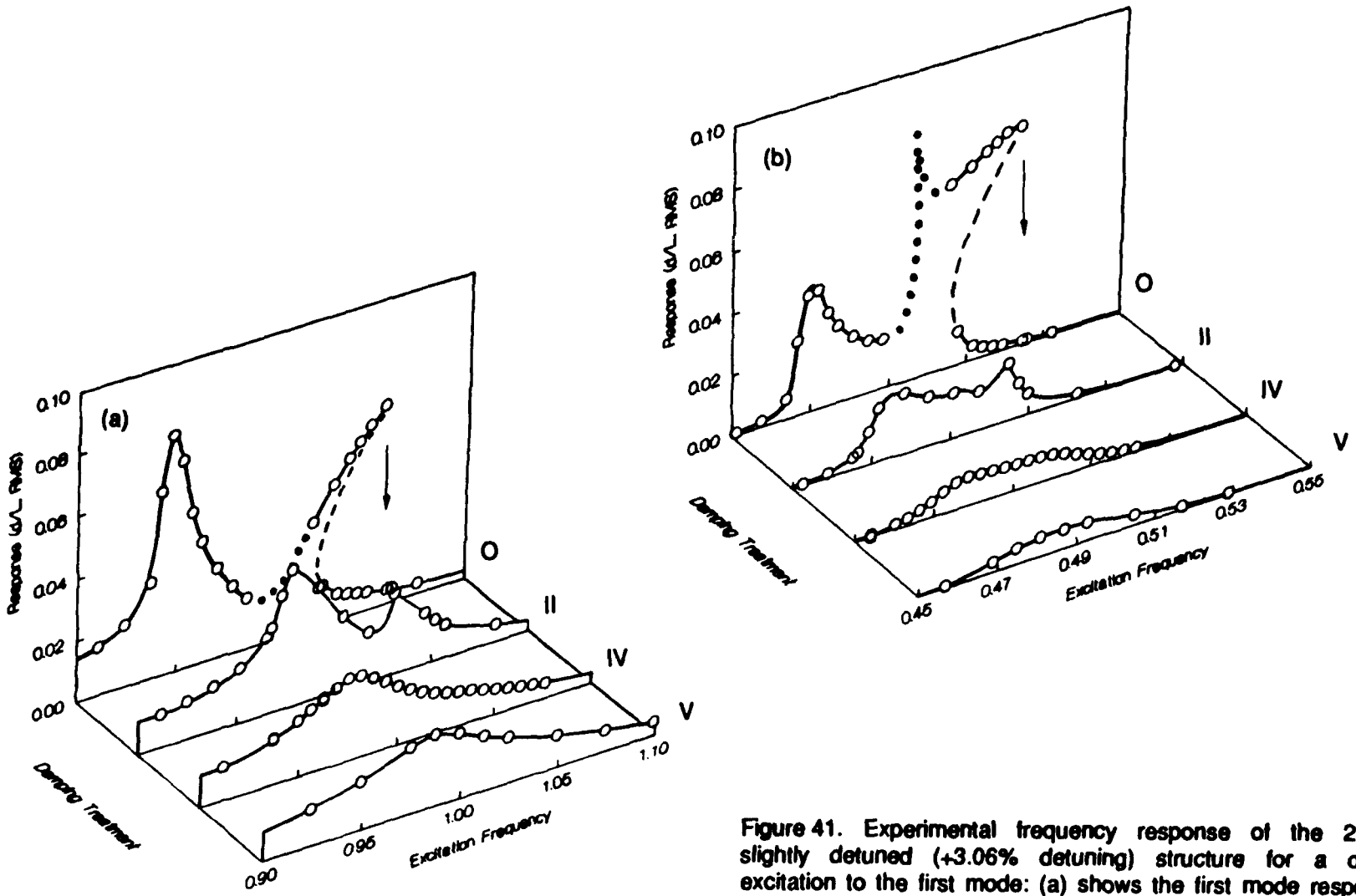


Figure 41. Experimental frequency response of the 2DOF slightly detuned (+3.06% detuning) structure for a direct excitation to the first mode: (a) shows the first mode response and (b) shows the second mode response. The dotted portion of the response curve corresponds to the modulation region where no steady state can be achieved. The arrows indicate jumps.

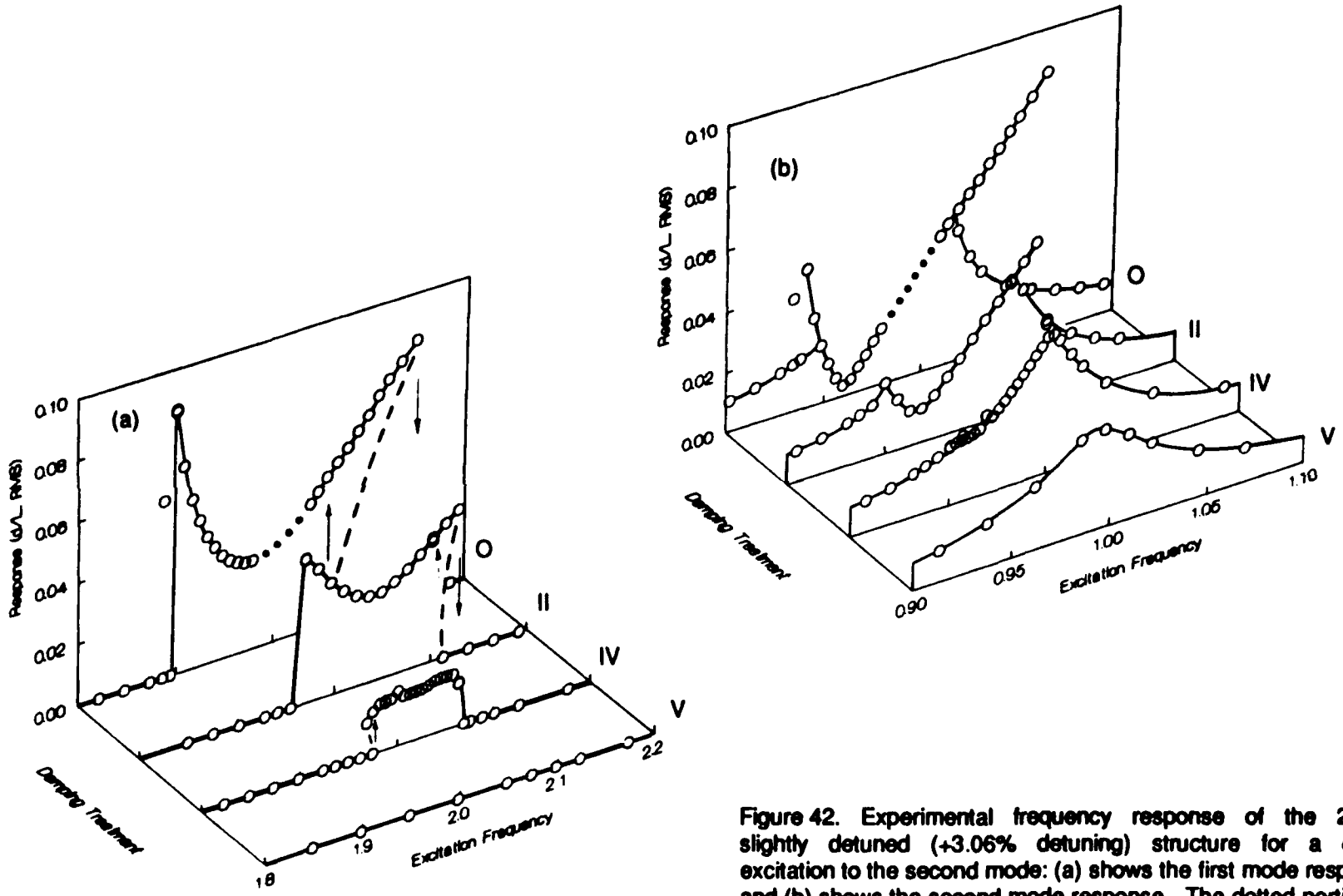


Figure 42. Experimental frequency response of the 2DOF slightly detuned (+3.06% detuning) structure for a direct excitation to the second mode: (a) shows the first mode response and (b) shows the second mode response. The dotted portion of the response curve corresponds to the modulation region where no steady state can be achieved. The arrows indicate jumps.

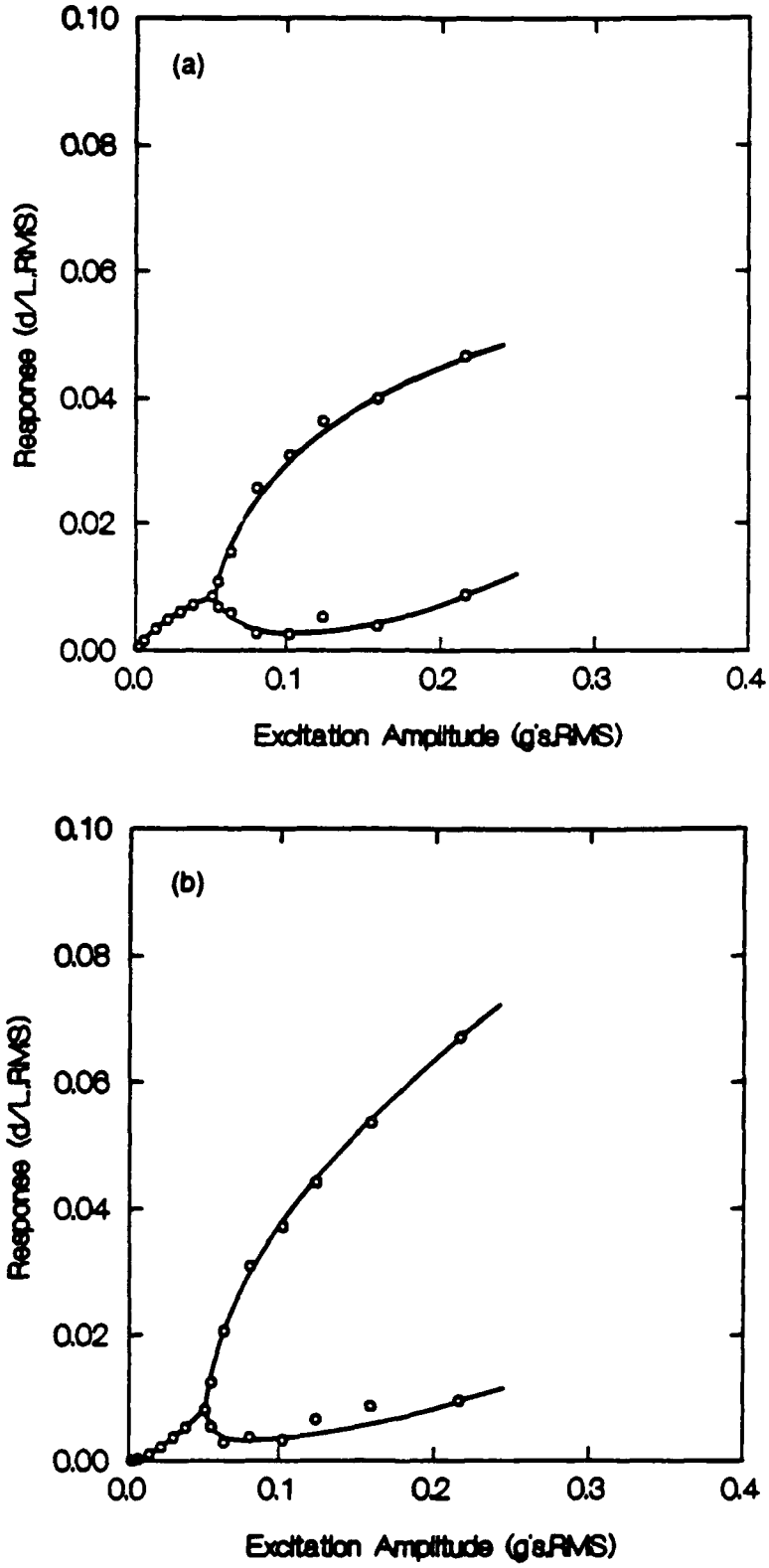


Figure 43. Experimental amplitude response of the 2DOF structure with +3.06% detuning and no damping treatment to a direct excitation to the first mode at a frequency of 0.9941 (for the first mode and 0.4922 for the second mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the modulation region in Figure 41. The curves approximate the bounds on the modulation.

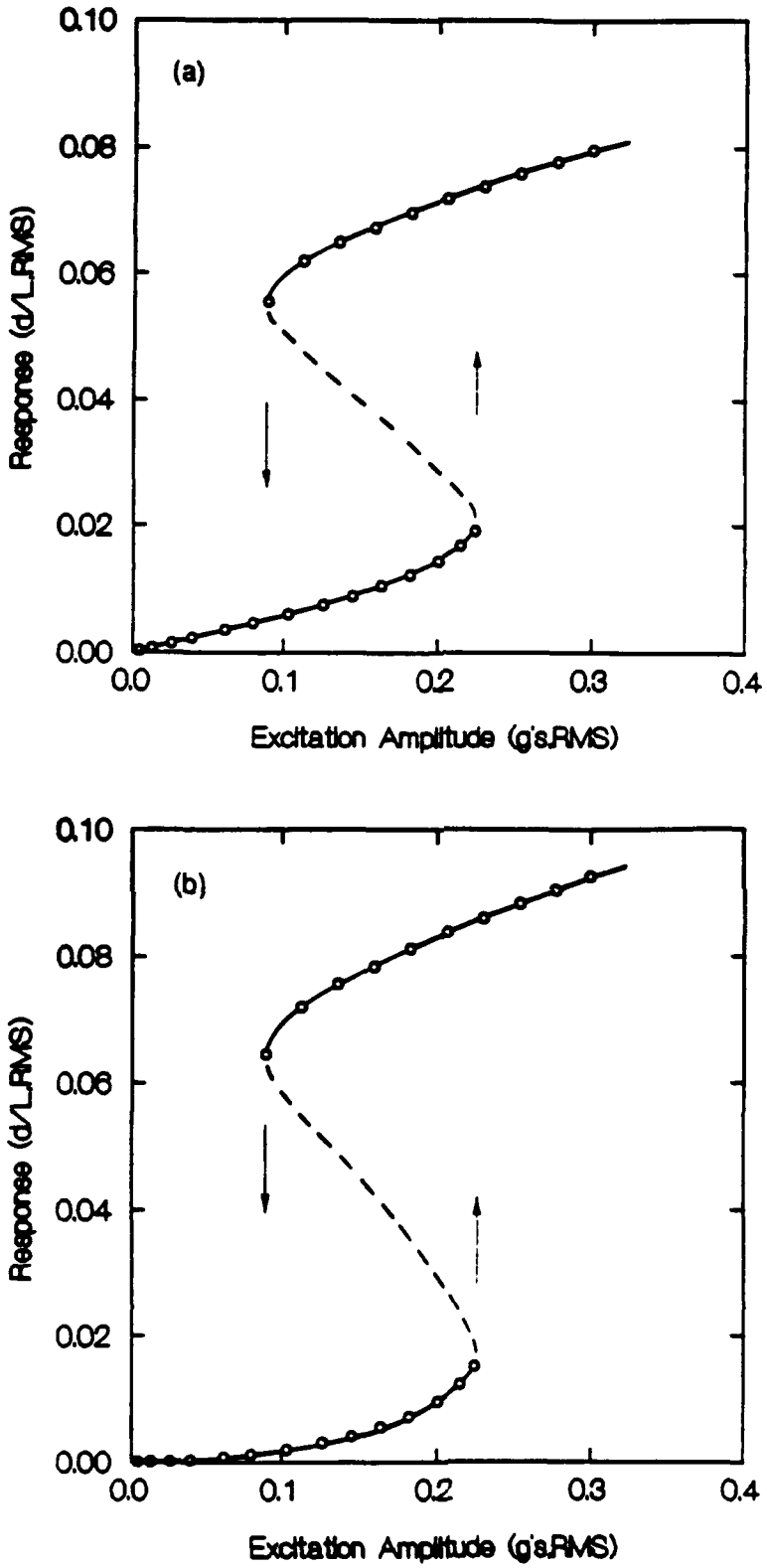
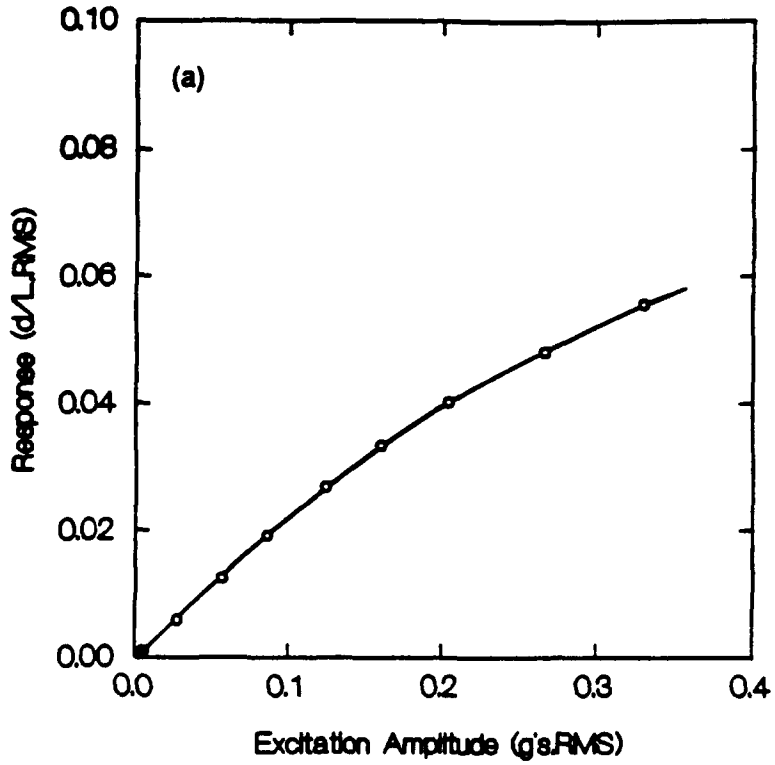


Figure 44. Experimental amplitude response of the 2DOF structure with +3.06% detuning and no damping treatment to a direct excitation to the first mode at a frequency of 1.0488 (for the first mode and 0.5165 for the second mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the overhang region in Figure 41.



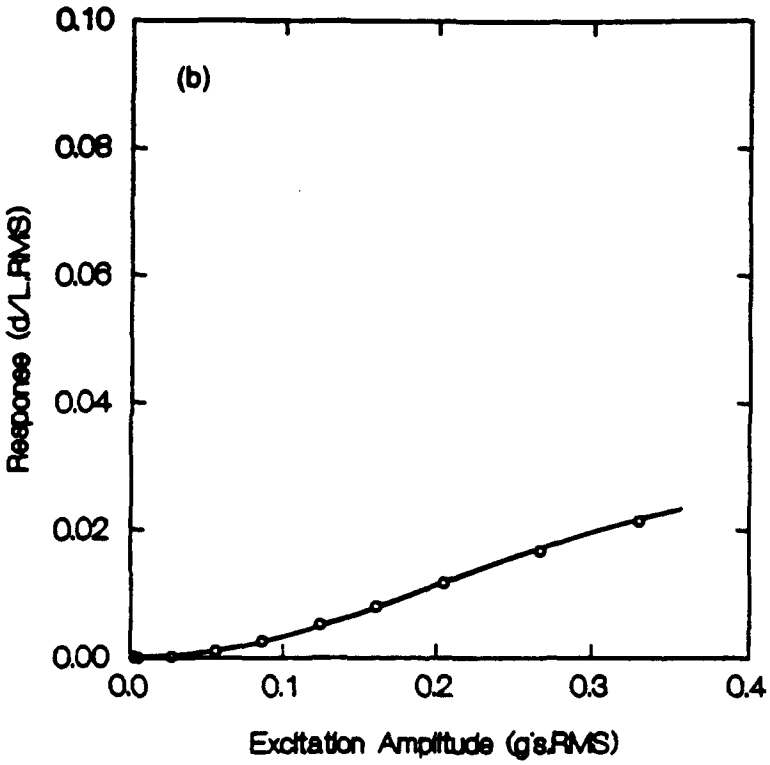


Figure 45. Experimental amplitude response of the 2DOF structure with +3.76% detuning and maximum damping treatment (case V)to a direct excitation to the first mode at a frequency of 0.9866 (for the first mode and 0.4747 for the second mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the resonance region in Figure 41 [V].

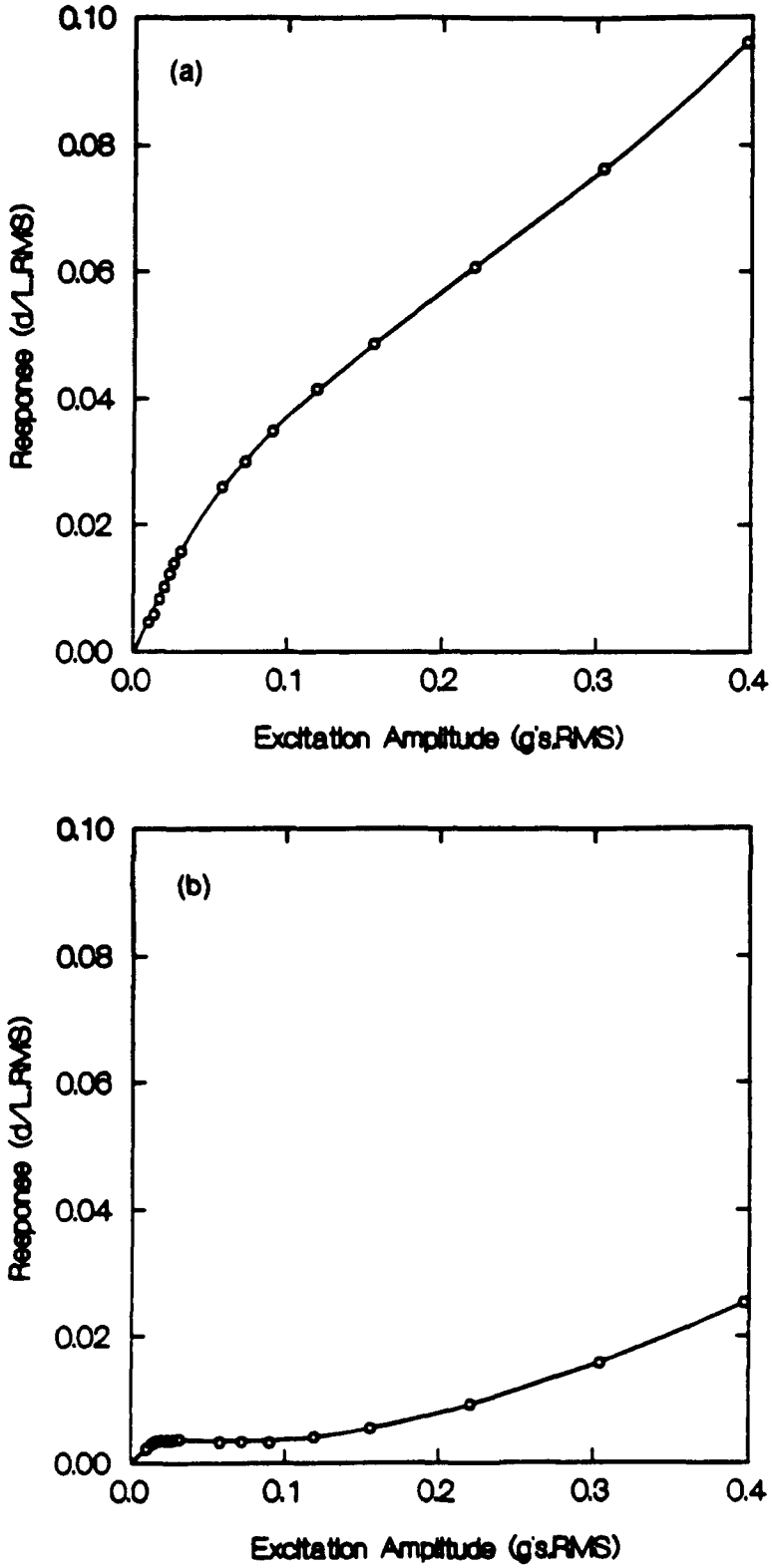
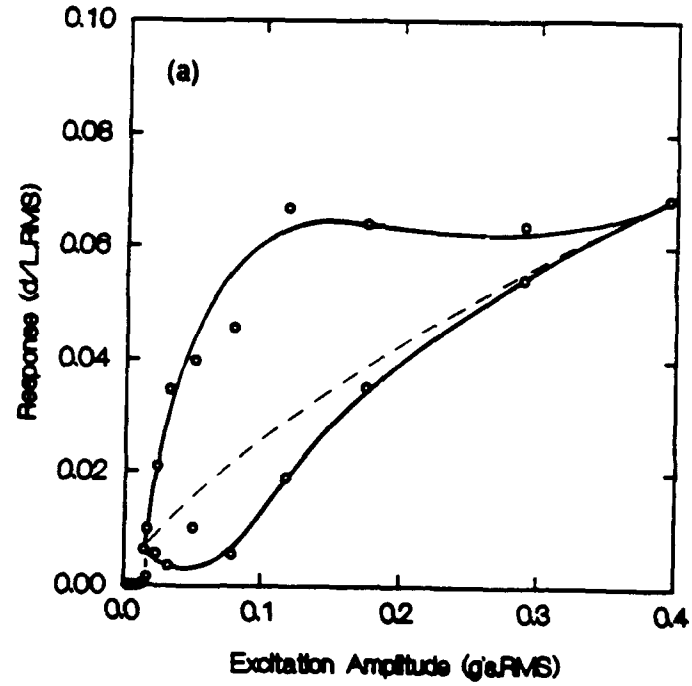


Figure 46. Experimental amplitude response of the 2DOF structure with +3.06% detuning and no damping treatment (case O) to a direct excitation to the second mode at a frequency of 0.9600 (for the second mode and 1.9388 for the first mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the region just left of the modulation region in figure 42 [O].



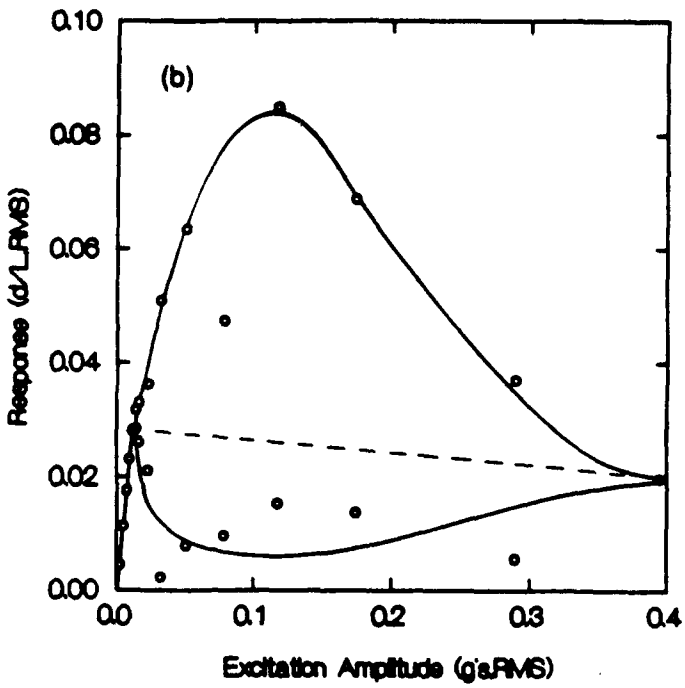
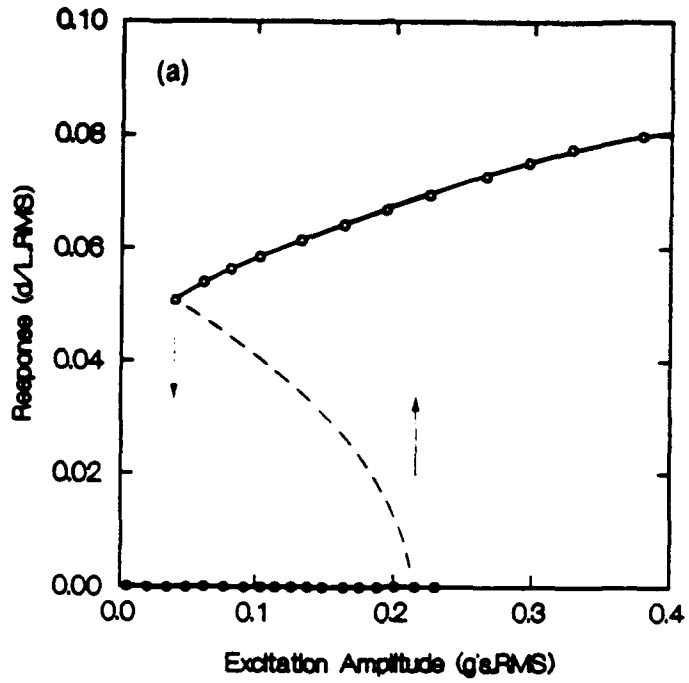


Figure 47. Experimental amplitude response of the 2DOF structure with +3.06% detuning and no damping treatment (case O) to a direct excitation to the second mode at a frequency of 0.9909 (for the second mode and 2.0012 for the first mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the modulation region in figure 42 [O]. Note that the modulation can be suppressed with sufficiently large amplitudes of excitation; however, with the large amplitude of excitation to the second mode we get a large response of the first mode because the second mode is essentially saturated.



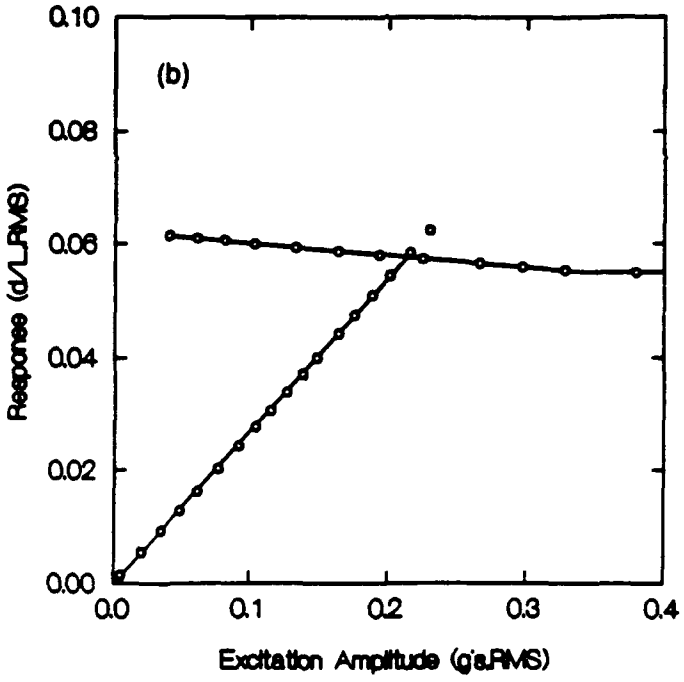


Figure 48. Experimental amplitude response of the 2DOF structure with +3.06% detuning and no damping treatment (case O) to a direct excitation to the second mode at a frequency of 1.0309 (for the second mode and 2.0819 for the first mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the overhang region to the right of the modulation region in figure 42 [O]. The arrows indicate jumps. The isolated point to the right of the bifurcation indicates the temporary stationary value achieved by the structure immediately after the amplitude of excitation was increased. However, since it is an unstable response, by waiting long enough the response went to the stable nonlinear response.

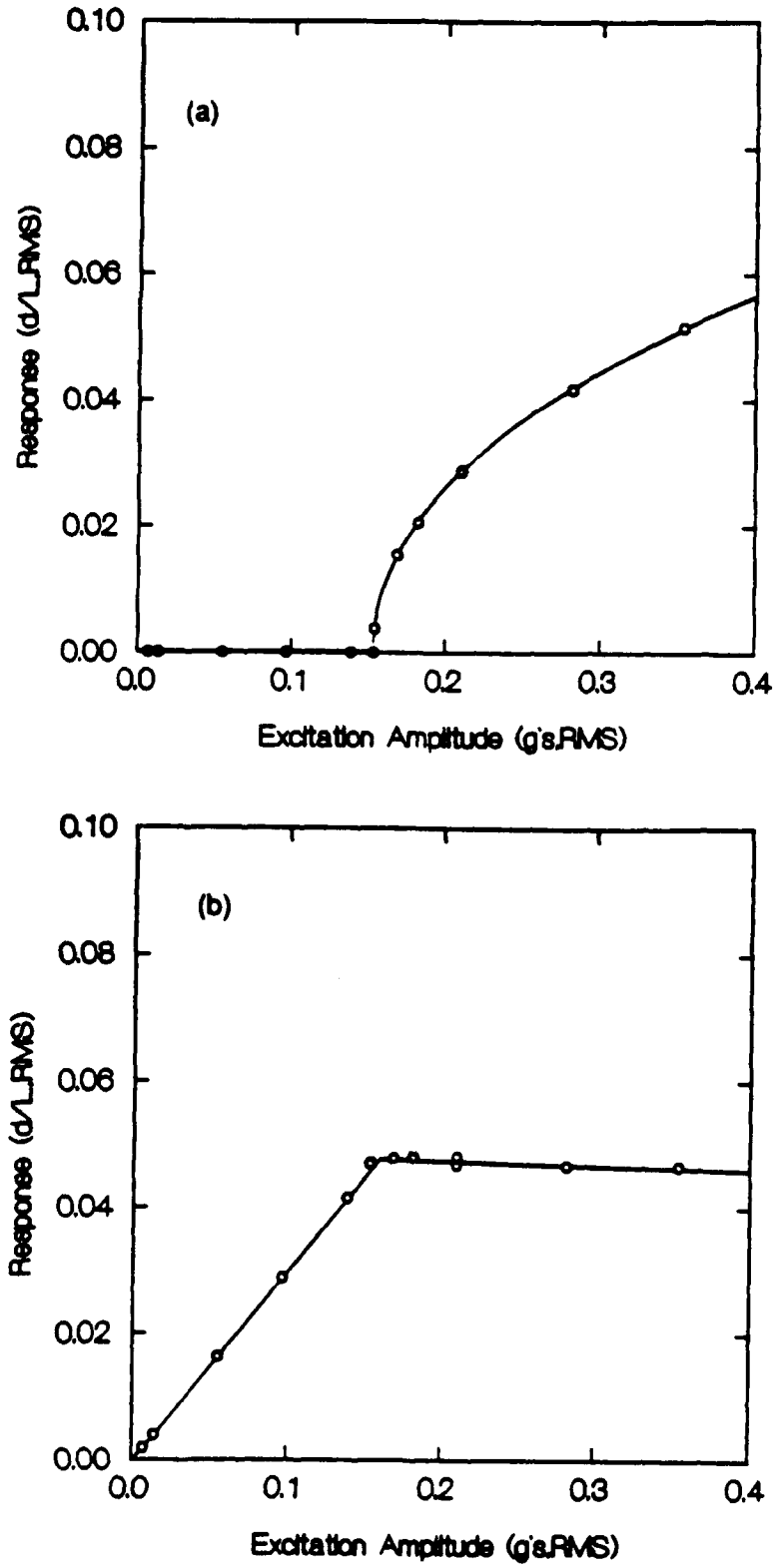


Figure 49. Experimental amplitude response of the 2DOF structure with +3.76% detuning and maximum damping treatment (case V) to a direct excitation to the second mode at a frequency of 1.0309 (for the second mode and 2.0819 for the first mode): (a) first mode amplitude, (b) second mode amplitude. This frequency corresponds to the peak in Figure 42 [V].

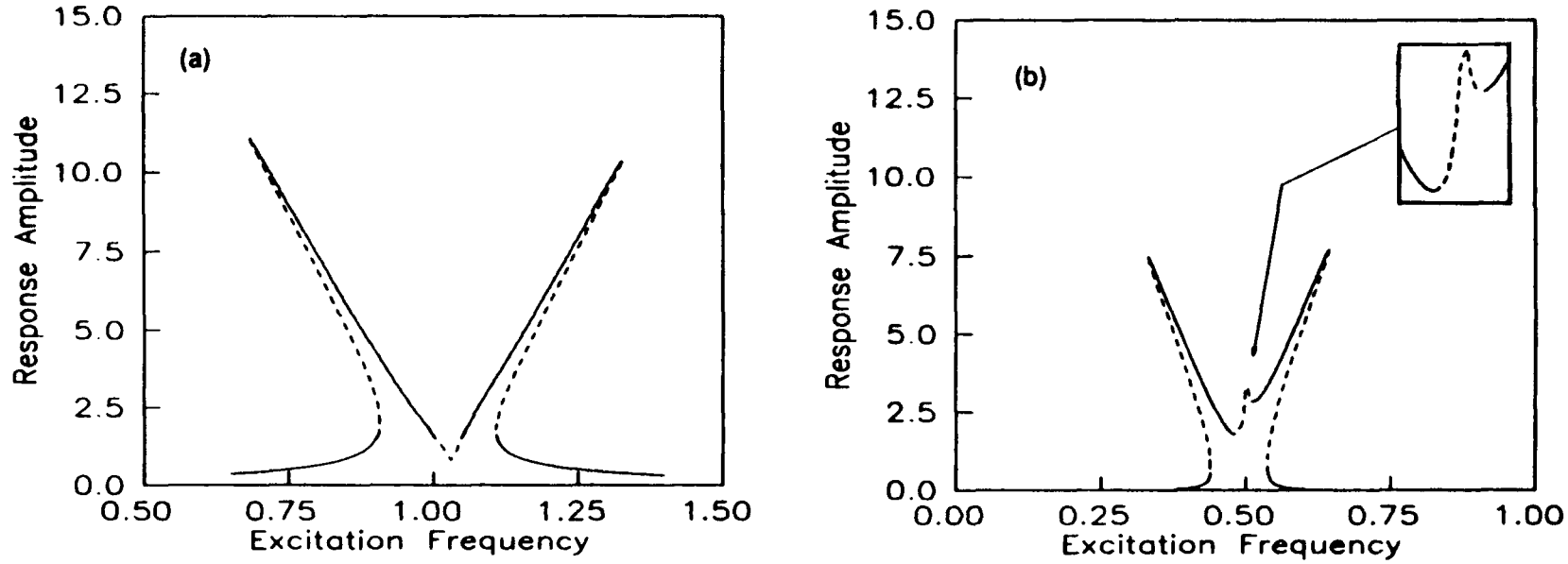


Figure 50. Theoretical frequency response for a direct excitation to the first mode (compare with Figure 41-[O]) for the slightly detuned structure before any damping treatment was applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The theory captures the qualitative features of the experimental responses. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 35. The dashed line represents unstable solutions; hence these could never be realized in an experiment.

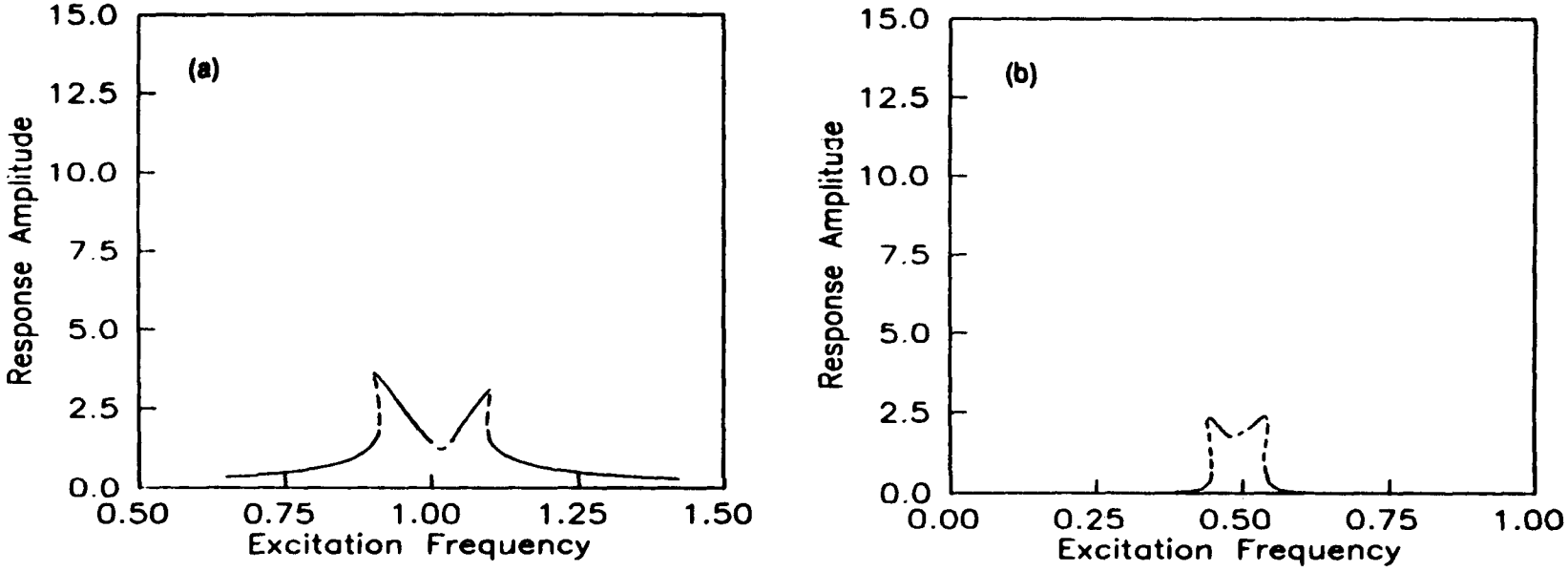


Figure 51. Theoretical frequency response for a direct excitation to the first mode for the slightly detuned structure for damping treatment III: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 38:

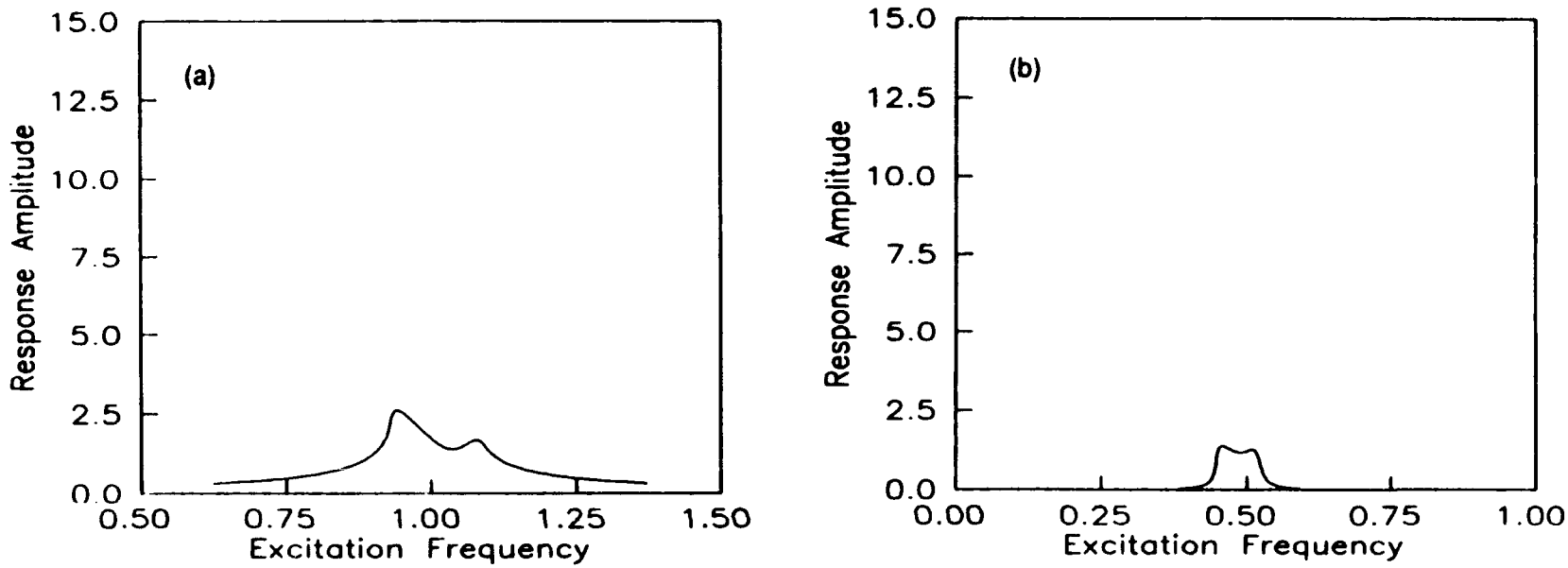


Figure 52. Theoretical frequency response for a direct excitation of the first mode (compare with Figure 41-[V]) for the slightly detuned structure for the maximum amount of damping treatment applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The estimated values of the damping coefficients and the natural frequencies were obtained from the complex exponential curve fit shown in Figure 40.

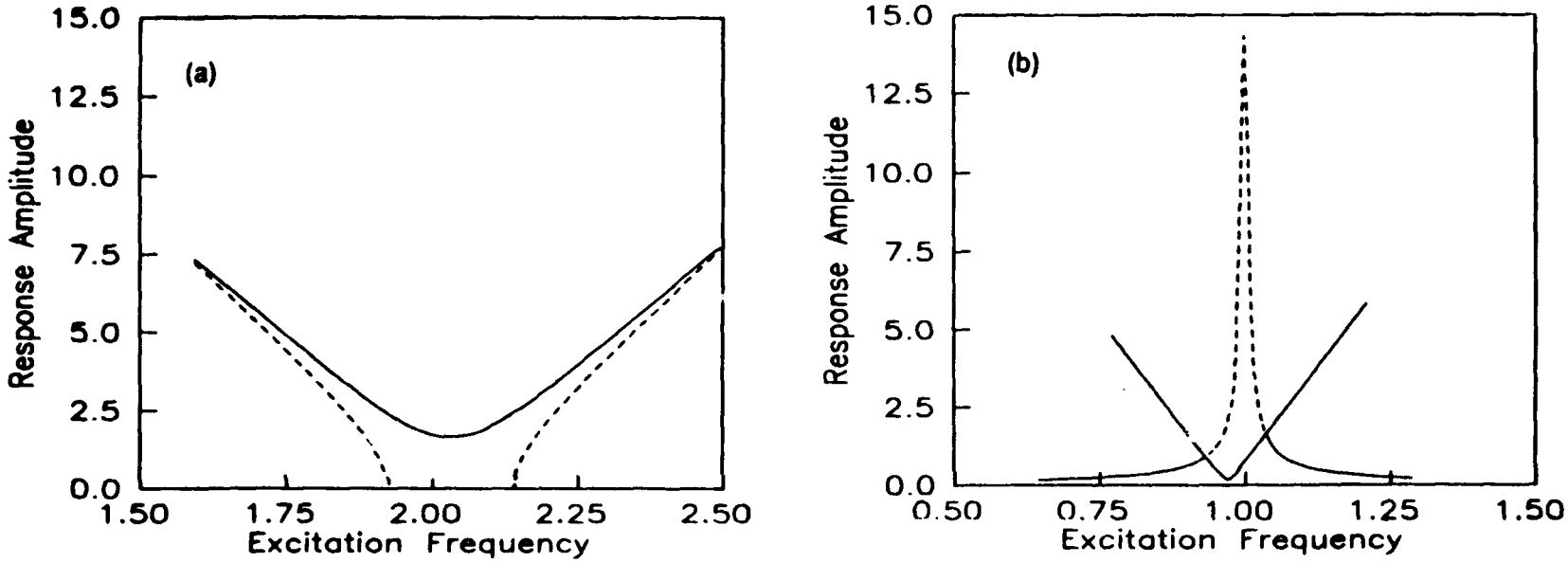


Figure 53. Theoretical frequency response for a direct excitation to the second mode (compare with Figure 42-[O]) for the slightly detuned structure before any damping treatment was applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The theory captures the qualitative features of the experimental responses. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 35. The dashed line represents unstable solutions; hence these could never be realized in an experiment.

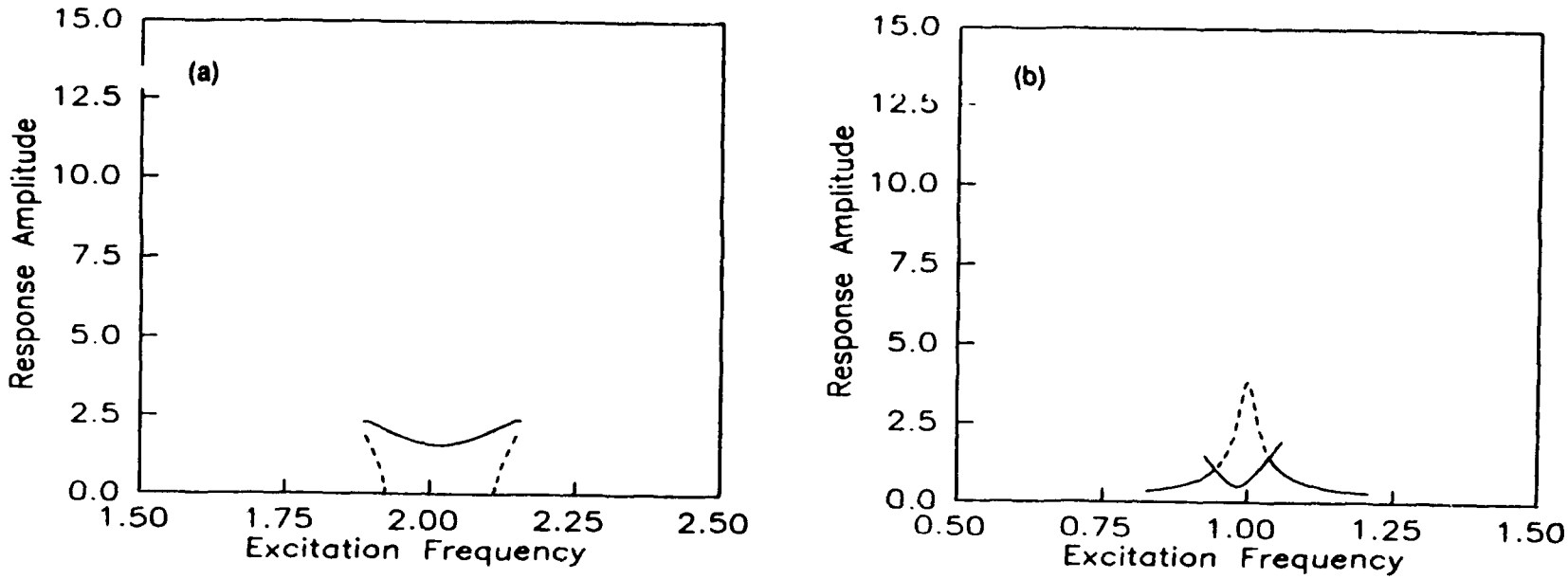


Figure 54. Theoretical frequency response for a direct excitation to the second mode for the slightly detuned structure for damping treatment III: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 38:

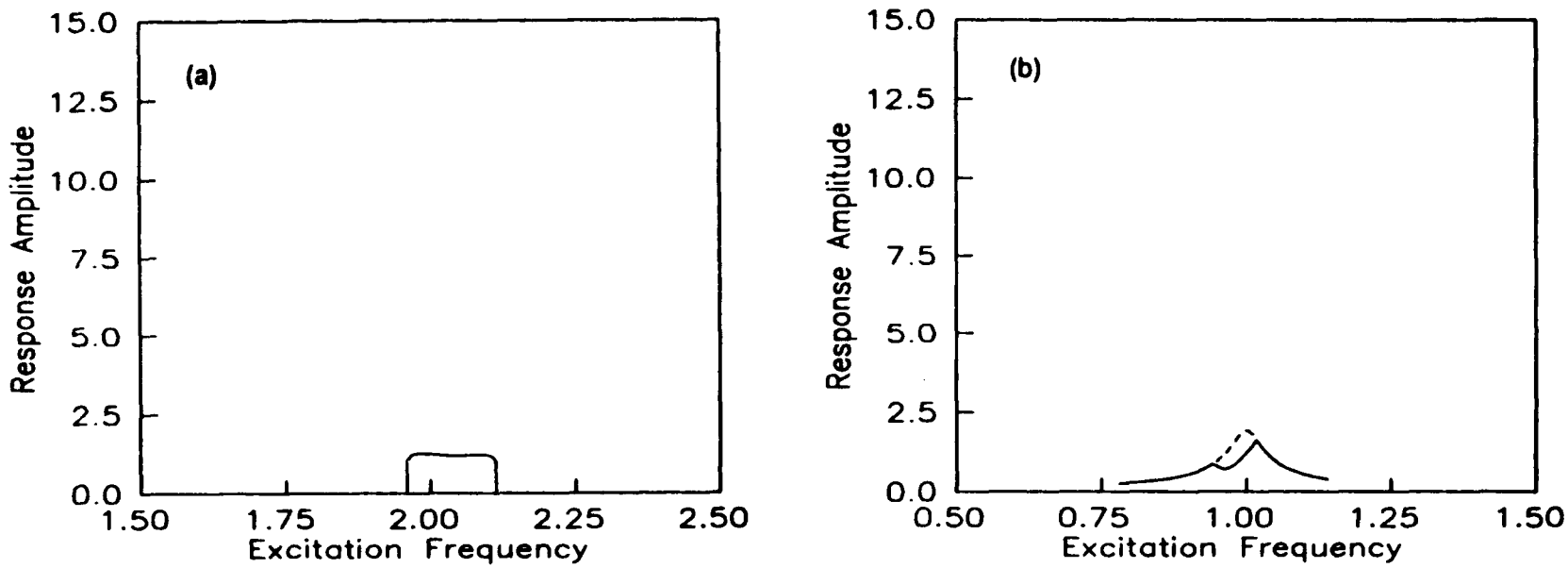


Figure 55. Theoretical frequency response for a direct excitation of the second mode (compare with Figure 42-[V]) for the slightly detuned structure for the maximum amount of damping treatment applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The estimated values of the damping coefficients and the natural frequencies were obtained from the complex exponential curve fit shown in Figure 40.

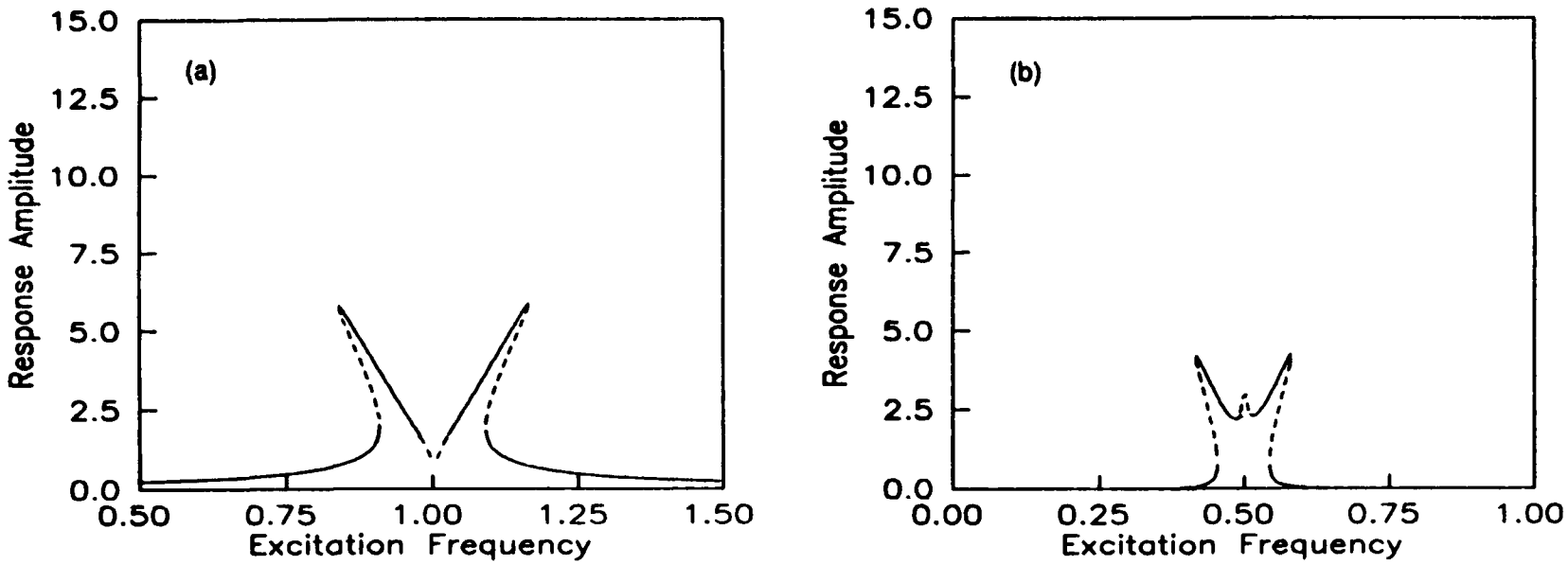


Figure 56. Theoretical frequency response for a direct excitation to the first mode (compare with Figure 27-[O]) for the tuned structure before any damping treatment was applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The theory captures the qualitative features of the experimental responses. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 22. The dashed line represents unstable solutions; hence these could never be realized in an experiment.

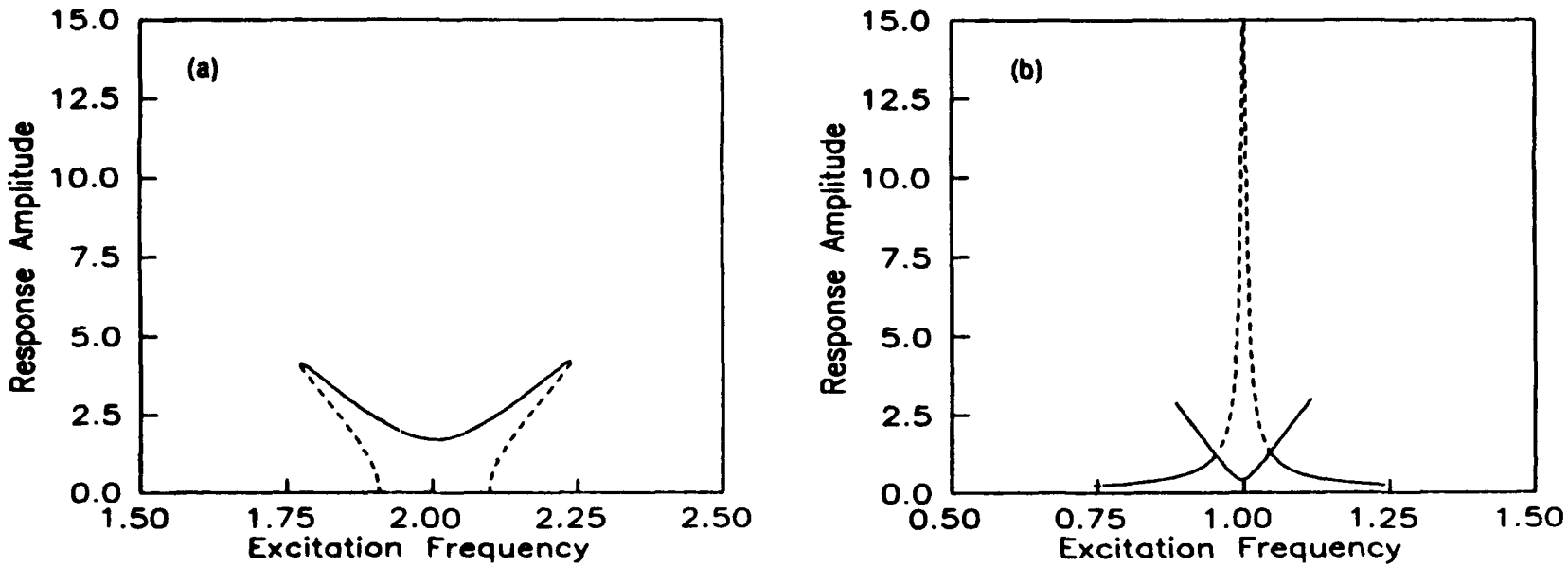


Figure 57. Theoretical frequency response for a direct excitation to the second mode (compare with Figure 28-[O]) for the tuned structure before any damping treatment was applied: (a) first mode response, (b) second mode response, (——) stable, (----) unstable. The values of the damping coefficients and natural frequencies were obtained from the complex exponential curve fit shown in Figure 22.

FINAL REPORT

UES PROJECT 210
S-210-11MG-002

TITLE

AM1 Molecular Orbital Calculations of the Conformational Properties of Odd-electron Rigid Rod Polymer Model Species

Submitted by John W. Connolly
Dept. of Chemistry
University of Missouri-Kansas City
Kansas City, MO 64110

Duration:
Nov. 1, 1990 to Oct. 31, 1991

AM1 Molecular Orbital Calculations of the Conformational Properties of Odd-electron Rigid Rod Polymer Model Species

John W. Connolly
Dept. of Chemistry, University of Missouri-Kansas City
Kansas City, MO, 64110

Abstract:

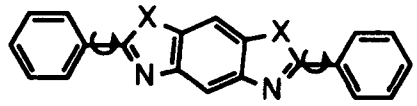
The summer 1990 SFRP.project is continued to include AM1 semi-empirical calculations on odd-electron rigid-rod polymer model species. The justification for the continued series of calculations is that evidence has been found for odd-electron species in bulk sample of rigid rod polymer fibers. Torsional barriers calculated for neutral radical species are about 3 Kcal/mol higher than corresponding closed-shell species. Radical cation and anion species have torsional barrier heights approximately four-times as great as the neutral closed- shell species. These effects are explained in terms of the p bonding characteristics of the frontier orbitals.

Introduction:

The rigid rod polymer, poly(p-phenylenebenzobisoxazole), PBO, and its sulfur analog, poly(p-phenylenebenzobisthiazole), PBT, have been found to exhibit exceptional specific strength and modulus, thermooxidative stability and environmental resistance when made into films and fibers[1]. The properties of these compounds are a consequence of the chain stiffness as well as the molecular stability due to extensive conjugation along the polymer chain.

At the molecular level, one aspect of chain stiffness is the barrier to rotation about the carbon-carbon single bonds linking the phenylene group to the aromatic heterocycle. Using a suitable model compound this barrier height can be calculated using semi-empirical molecular orbital techniques by calculation of the molecular heat of formation at each of a series of molecular conformations. Figure 1 shows two model compounds on which such calculations have been recently reported[2,3].In both cases heats of formation were calculated at 10° intervals of rotation about the indicated carbon-carbon bonds. It should be noted that in this figure the symbol X stands for O, NH, or S.

A



B

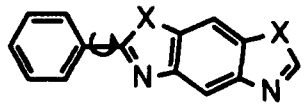


FIG 1. Model Compounds Used in Previous AM1 Calculations.

Figure 2 shows the model compound, again with X symbolizing O, NH, or S, on which an extensive series of AM1 calculations were done during the 1990 SFRP project.

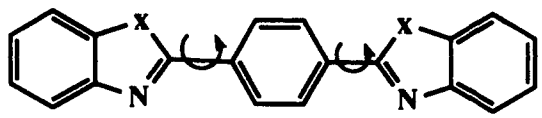


FIG 2. Model Compounds Structure Used in Summer 1990 SFRP.

In the previous calculations the model compounds started in the planar conformation and one segment of the molecule was rotated relative to the original plane of the molecule, as indicated in the arrows in the figures. In addition to that type of calculation we have examined the situation in which the one heterocyclic system is perpendicular to the remainder of the molecular plane. This structure is shown in Figure 3.

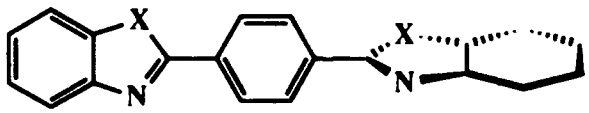


FIG 3. Perpendicular Model Compound

Since dynamics calculations on PBO and PBZT model structures[4] indicate that these materials are more flexible than the calculated torsional barrier

heights suggest, we decided to examine torsional barriers which model segments of the polymer chain which are perpendicular to one another.

Figure 4 shows the result of our AM1 calculations on the unsubstituted PBO model compound.

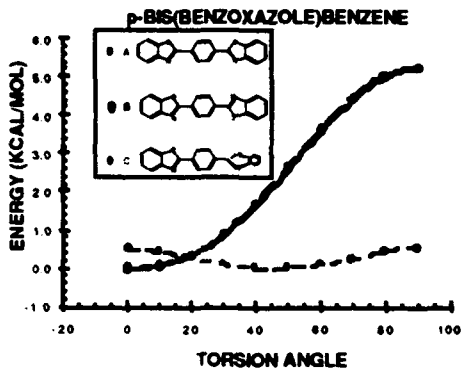


Fig 4. Graphical Display of AM1 Calculations on PBO Model Compounds.

Note the very small rotational barrier shown in curve C. The relevance.of this result is as follows. When a given polymer link rotates 90° it becomes perpendicular to adjacent links in the polymer. According to our calculations, the links adjacent to the perpendicular one then have much greater torsional flexibility. Table I lists all the molecular species on which calculations were carried out and Table II summarizes the results of these calculations.

TABLE I

STRUCTURAL KEY FOR TABLE II

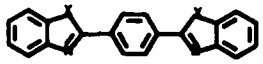
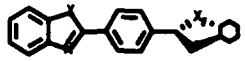
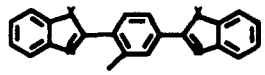
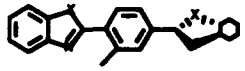
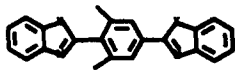
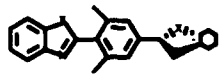
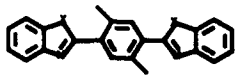
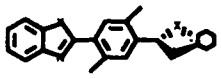
SYMBOL	STRUCTURE
I	
II	
III	
IV	
V	
VI	
VII	
VIII	

Table II summarizes the results of these calculations.

TABLE II

NUMERICAL SUMMARY OF AM1 CALCULATIONS

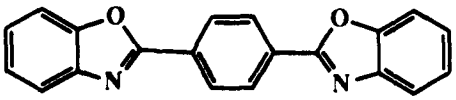
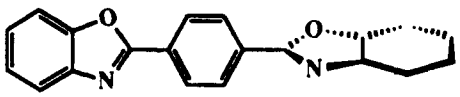
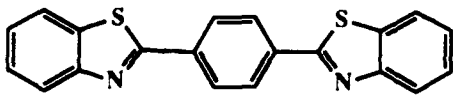
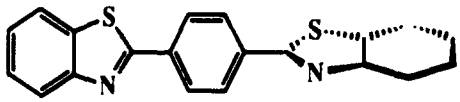
STRUCTURAL TYPE	TORSIONAL ANGLE (DEG) MIN ENERGY X=O	X=NH	X=S	TORSIONAL ANGLE (DEG) MAX ENERGY X=O	X=NH	X=S	BARRIER (KCAL/MOL) X=O	X=NH	X=S
I	0	30	10	90	90	90	5.2	2.7	2.2
II	45	45	45	0, 90	0, 90	0, 90	0.5	1.3	0.5
III	0	140	30	90	180	90	4	3.1	1.3
IV	20	150	60	90	0	0	1.7	4	1.6
V	40	50	70	90	0	0	1.7	7.1	1.7
VI	20	50	30	90	0	0	2.5	4.7	0.2
VII	80	50	90	90	0	0	4.2	8.5	3.7
VIII	140	130	45	0	180	0	1	3.1	1.1

In the current project we have continued these calculations to include charged and uncharged odd-electron model species. The significance of these new calculations relates to the chemical consequences of the processessing of synthetic PBO and PBZT. It has been reported[5] that the mechanical stress involved in the processing causes radical formation, as indicated by the presence of ESR signals in the processed polymer films. We then consider whether such chemical alteration of the polymer would cause a change in its stiffness.

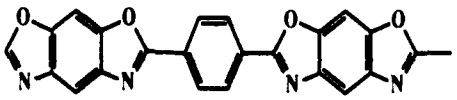
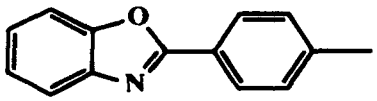
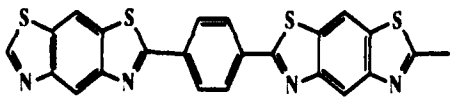
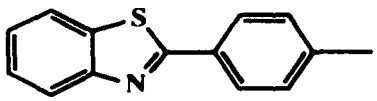
In Table III are summarized the results of our calculations on the closed-shell, cation radical, and anion radical PBO and PBZT model species.

TABLE III

SUMMARY OF AMI CALCULATIONS ON MODEL PBO AND PBZT SPECIES

COMPOUND	BARRIER HEIGHT (KCAL/MOL)		
	CLOSED SHELL	CATION RADICAL	ANION RADICAL
	5.2	18.8	21.8
	0.5	4.2	0.3
	2.2	11.8	18.5
	0.3	0	1.6

NEUTRAL RADICAL SPECIES

	PLANAR	PERPENDICULAR
	8	0
	8	
	5	0.3
	5	

The substantial increase in barrier height for both the cation radicals and the anion radicals can be understood in terms of the frontier orbitals for these species. The HOMO is shown if Figure IV. This orbital is

HOMO

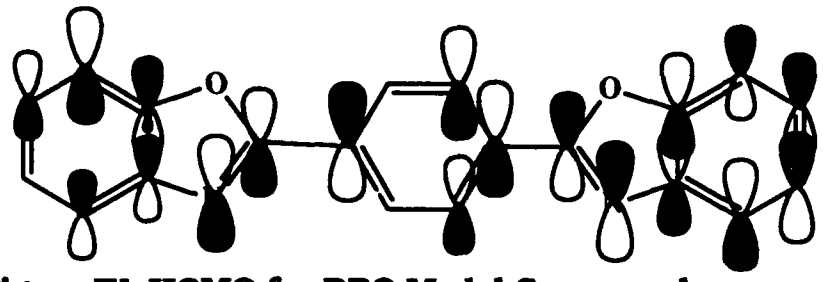


Figure IV. HOMO for DBO Model Compound.

antibonding across the sigma-bond about which rotation is occurring. When an electron is removed from this orbital (to make the cation radical) the repulsion across the sigma bonds decreases and rotation about these bonds is more difficult. The LUMO is shown in Figure V. This orbital is P bonding

LUMO

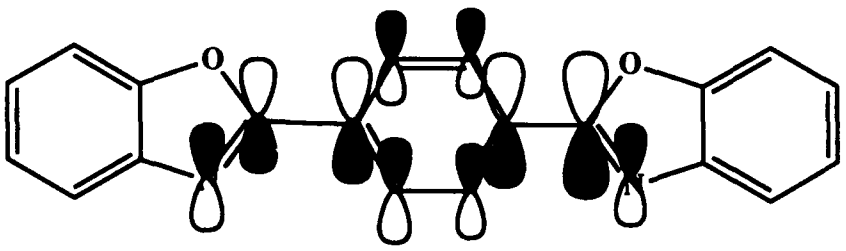


Figure V. LUMO for DBO Model Compound.

across the C-C s bond. Partial occupancy of it, as occurs in the radical anion, increases the difficulty in rotation about this sigma bond.

It was necessary to change our model species to model neutral radical species in the polymer chain. Neutral radicals formed during processing of the polymer would presumably be formed by breaking a C-C s bond along the backbond. Such radicals would be carbon-centered., with the carbon being either at the end of the phenyl group or the heterocyclic group. The structures shown in Table III model both of these possibilities for PBO and PBZT. As can be seen by the data in Table III the barrier to rotation in the flat structures is about 3 Kcal/mol higher than in closed-shell species. In the perpendicular structure the barrier to rotation is essentially zero.

In summary, our calculations suggest that if the processessing of these rigid rod polymers causes radical formation. The resulting polymer chain

segments will be less flexible, even though the radical formation could cause the chain length to be reduced.

REFERENCES:

1) Evers, R. C., Arnold, F. E. and Helminiak, T. E., **Macromolecules**, (1981), 14, 925

2) Yang, Y. and Welsh, W. J., **Ibid.**, (1990), 23, 2410

3) Farmer, B. L., Wiershke, S. G., and Adams, W. W., **POLYMER**, (1990), 31, 1637

4) Farmer, B. L., Wiershke, S. G. **Ibid**, in press

5) VanderHart, D. L., Wang, F. W., Eby, R. W., Fanconi, B. M. and DeVries, K. L., "Exploration of Advanced Characterization Techniques for Molecular Composites", AFWAL-TR-85-4137, February 1986

1990 USAF-UES RESEARCH INCENTIVE PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

STRUCTURAL ANALYSIS OF POTENTIAL NLO CHROMAPHORES

Prepared by:	David A. Grossie, Ph.D.
Academic Rank:	Assistant Professor
Department and	Department of Chemistry
University:	Wright State University

Date:	31 Dec 91
Contract No:	F49620-88-C-0053/SB5881-0378

ACKNOWLEDGEMENTS

I would like acknowledge the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. Additionally, acknowledgement of Universal Energy Systems must be given for their assistance in the administrative aspects of the program.

I wish to thank the Polymer Branch of the Materials Laboratory for access to the Enraf-Nonius CAD4 single-crystal diffraction system, on which all data was collected.

STRUCTURAL ANALYSIS OF POTENTIAL NLO CHROMAPHORES

by

David A. Grossie, Ph.D.

ABSTRACT

Crystallization studies on a collection of benzothiazole and tetrazine derivatives were conducted, with the intent to produce crystals suitable for single-crystal X-ray diffraction. The benzothiazole compounds are designed to model a single repeating unit of a polymer chain, with an electron-withdrawing group available for attachment of a pendent having potential nonlinear optical (NLO) properties. These compounds were also examined via molecular mechanics to determine any intra-molecular interactions which may affect the molecular packing in the crystal lattice. Single-crystal x-ray diffraction data was collected on Bis(phenyl)tetrazine and the molecular structure determined. The tetrazine derivative, a molecular analogue of para-terphenylene, crystallizes in an monoclinic crystal lattice with cell constants of a=5.448(2), b=5.119(1), c=20.236(7) Å and β = 93.64(2)^{o}. The observed space group is $P2_1/n$, a centric space group. The structure was solved and refined, yielding a R-factor of 0.094. The compound is planar with little distortion in the internal bond distances and angles.

I. INTRODUCTION:

The Polymer Branch of the Materials Laboratory at the Wright Aeronautical Laboratory, Wright-Patterson Air Force Base and the Department of Chemistry at Wright State University are interested in the synthesis and characterization of polymeric materials. Basic research is also conducted in the structure of polymeric materials and the correlation of the structure and physical properties. The emphasis of this area is to predict the properties of a polymer prior to its synthesis. In this way, the synthesis problem can have greater direction and produce new and better materials with more efficiency. One of the techniques used in determining the structure of polymers is to examine by single-crystal X-ray diffraction methods compounds that may be used to form the backbone, pendants, or cross-links of the polymer. By knowing the structure of a small, repeating portion of the polymer, the polymer itself may be mathematically modeled, yielding the physical properties. The synthesis of a new material does not always produce a crystalline, solid compound, so crystallization studies are used to find the optimal conditions under which good crystals may be grown. Once the crystalline material has been produced and a structure determination completed, examination of the molecule via molecular mechanics will isolate the intra- and inter-molecular effects in the packing of the molecule in the crystal lattice.

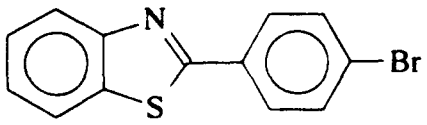
II. OBJECTIVES OF THE RESEARCH EFFORT:

A study of model compounds of polymeric materials that have potential nonlinear optical properties will be conducted. This study will involve the structural analysis of compounds by single-crystal x-ray diffraction techniques, with the intent to amass data which may be used to correlate the observed structure and the magnitude of the nonlinear optical response. The primary structural information that is needed by the currently accepted theories is the centricity of the crystal lattice in which the compound of interest crystallizes and the extent of π-orbital conjugation. To this end, the study will examine compounds with known nonlinear optical properties, and other conjugated systems with separated electron donor and acceptor groups. In addition, the analysis will be conducted into the means of attaching the NLO active compound to a polymer chain through the examination of polymeric models with available electron donor or acceptor functionalities.

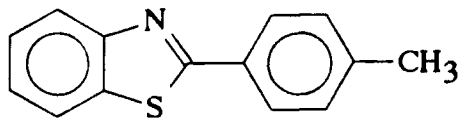
III.

As part of the research effort, attempts were made to grow crystals of various benzothiazole, thiophene and tetrazine derivatives. In most cases, a small portion of the compound was placed in a test tube, a small amount of the solvent added, and the mixture heated to dissolve the solid material. In a few cases, attempts were made to co-crystallize the compound with another compound that may (via dipole interactions) help orient the molecules and enhance the crystallinity of the recrystallized material. The results of the study are tabulated below, followed by the structural formulas of the compounds studied.

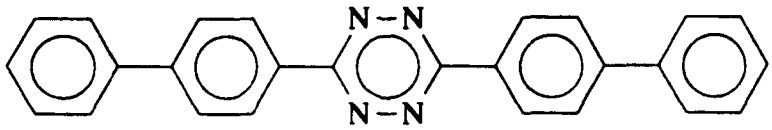
Compound	**Solvents**	**Results**
#51	Ethanol	long, thin crystals
#53	Ethyl acetate	large, thin platelets
#54	Toluene DMF	amorphous powder amorphous powder
Thio #1	Ethanol Ethanol/toluene-TPO co-crystallization Ethanol-naphthol co-crystallization Ethanol-phenol co-crystallization	amorphous powder colorless long, thin crystals no results no results
Thio #2	Ethyl acetate	small, thin crystals
Tet #1	Ethyl acetate	large red crystals
Tet #2	Not soluble	



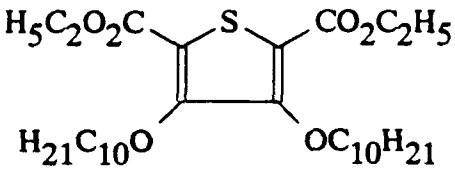
#51



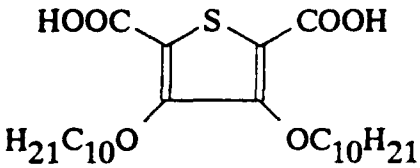
#53



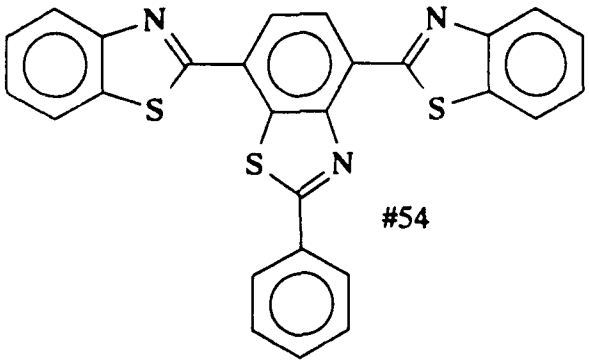
Tet #2



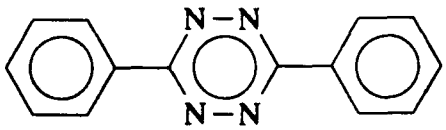
Thio #1



Thio #2



#54



Tet #1

IV.

a. Crystalline samples of a series of compounds from the crystallization study were examined using an optical microscope to determine the size and quality of the individual crystals. One of the compounds examined showed promise of containing suitable crystals for diffraction analysis, whereas the remainder were of insufficient size and quality to be analyzed. A single crystal of para-Bis(phenyl)tetrazine (Tet #1) were prepared for analysis by attaching it to a thin glass fiber and placing it at the center of an Enraf-Nonius CAD-4 automated diffractometer. Preliminary x-ray analysis of the selected crystal was made. These results are summarized in Table 1, along with the parameters of the subsequent data collection.

b. Data collected on the crystalline sample was examined for the presence of space-group-determining systematic absences using the program LOOK (Chapius, 1984). For the sample, an appropriate space group was determined--$P2_1/n$.

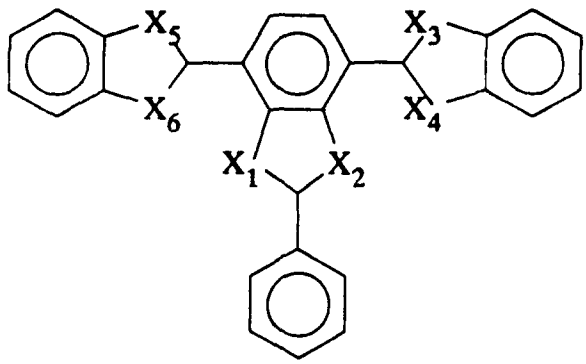
c. Using the direct methods routines contained in the XTAL system of crystallographic programs (Grossie, 1991; Hall and Stewart, 1990), the initial structures were determined. These structures were refined using CRYLSQ (Olthof-Hazekamp, 1990) a full-matrix least-squares refinement program contained in XTAL (Hall and Stewart, 1990).

d. Figure 1 shows an ORTEP (Johnson, 1971; Davenport, Hall, and Dreissig, 1990) drawing of the refined structure of p-Bis(phenyl)tetrazine. A listing of the interatomic distances and angles is presented in Table 2. The molecule is completely planar, with a maximum deviation from planarity of 0.011(4) Å.

e. The molecule lies on an inversion center such that each atom at one end is related to a second at the opposite end. Thus the attached phenyl rings are absolutely identical. The phenyl show typical distortions with bond distances ranging from 1.366 to 1.402 Å with an average of 1.387 Å. The bond angles show similar distortions with values ranging from 118.2 to 121.5° with an average of 120°. The tetrazine ring is slightly distorted with a widening of the bond angle at C (123.2°) and a narrowing at the two N atoms (117.9 and 118.9°).

V.

As a prelude to the X-ray crystal structure analysis, possible molecular structures for compound 54 were examined by molecular mechanics. The results for the four possible conformers are tabulated below. Of interest in this compound is the minima observed when changing the position of the sulfur and nitrogen atoms in the thiazole rings. As the positions are changed, there is a change in the molecular planarity and the minimum energy. In experimental determinations of melting point for this compound, three distinct melting points are observed. It is obvious then that three or perhaps all four of these conformers exist, with the highest melting point corresponding to the lowest molecular mechanical energy.



	X_1	X_2	X_3	X_4	X_5	X_6	Distance X_1 - X_6	Distance X_2 - X_4	Torsion X_4(°)	Torsion X_6(°)	MMX Energy
1	S	N	N	S	N	S	3.275	3.045	1.81	24.64	83.72
2	S	N	S	N	N	S	3.295	2.933	21.45	-20.27	85.89
3	S	N	S	N	S	N	3.126	2.889	3.87	16.02	85.49
4	S	N	N	S	S	N	3.058	3.009	3.37	14.16	82.95

VI. RECOMMENDATIONS:

There is a need to improve the crystallization techniques used in recrystallizing these compounds. The crystallization study needs to be broadened by the use of other methods besides solution recrystallization. Co-crystallization methods need to be more fully exploited, using a wide range of compounds for the co-crystallization. Co-crystallization has a great deal of potential for producing crystalline material of compounds that have low melting points (close to room temperature) or low solubility. This technique also provide a means by which a compound that crystallizes in a centrosymmetric lattice may be crystallized in an acentric lattice, thus allowing the second-order NLO properties of the compounds to be examined and used.

Additional crystal structures are needed to further quantify the prediction of NLO properties and to provide initial information for numerical calculations on known NLO active materials.

Currently, there are three factors that are assumed to produce the desired NLO response, with one factor being quantitative and a second based on a relative scale. These two factors are the centricity of the crystal lattice and the electron-donating and withdrawing effects of the commonly used functional groups. The third factor, the extent of conjugation within the molecule, is currently unquantified. Molecular planarity is normally taken as the first clue that a non-fused, π-bonded ring system is conjugated.

Since this information, like the centricity, is directly obtainable from the structural analysis of a crystalline compound, the extent of conjugation can be quantified by this process.

The molecular packing of a molecule within a crystal lattice is of great importance to the observation of NLO properties in the bulk material. One factor that influences molecular packing is the shape of molecule and the intra-molecular interactions. These interactions can be modelled easily using molecular mechanics, work that should be continued. The inter-molecular attractions also influence the packing. The inter-molecular interactions can be controlled using guest compounds that interfere with the attractive sites within a molecule that cause then to come together in the solid state. To study these effects, the co-crystallization of the potential NLO active material and an appropriate guest is recommended.

With the above pieces of data obtained and analyzed, the synthesis of nonlinear optical materials can be by rational design. This will allow the physical properties of the material to be optimized without compromising the desired nonlinear optical properties.

REFERENCES

Chapius, G., "LOOK. A FORTRAN Program for Generating Simulated Precession Photographs from Diffractometer Data," University of Lausanne, Switzerland, 1984.

Davenport, G., Hall, S. R., and Dreissig, W., (1990) "ORTEP" *XTAL 3.0 User's Manual*, Hall, S. R. and Stewart, J. M., Eds., Universities of Western Australia and Maryland.

Grossie, D., "Desktop Crystallography: XTAL on an Enhanced Personal Computer," American Crystallographic Association Meeting, 1991.

Hall, S. R. and Stewart, J. M., (1990) Eds., *XTAL 3.0 User's Manual*, Universities of Western Australia and Maryland.

Olthof-Hazekamp, R.,(1990) "CRYLSQ" *XTAL 3.0 User's Manual*, Hall, S. R. and Stewart, J. M., Eds., Universities of Western Australia and Maryland.

Johnson, C.K., ORTEP II. Report ORNL-3794, revised. Oak Ridge National Laboratory, Tennessee, 1971.

Table 1. Experimental Details for p-Bis(phenyl)tetrazine.

Formula:	$C_{14}H_{10}N_4$
Formula weight:	234.2
F(000):	244
Crystal dimensions:	0.20 x 0.41 x 0.62 mm
Color:	Red
Radiation:	Mo Kα
Wavelength:	0.71073 Å
Temperature:	23°
Crystal form:	monoclinic
Space group:	$P2_1/n$
Cell constants:	a = 5.448(2) Å
	b = 5.119(1) Å
	c = 20.236(7) Å
	β = 93.64(2)°
Volume:	563.2(3) Å^3
Z:	2
Density:	1.38 g/cm^3
Absorption coefficient:	0.95 cm^{-1}
Scan type:	ω/2θ
Scan rate:	1.20 - 5.58° /min
Scan width:	1.00 + 0.344 tan θ
Maximum 2θ :	70.0°
Reflections measured:	2827 total
	2461 unique
Corrections:	Lorentz-polarization
	Numerical absorption (range of T: 0.9619-0.9828)
	Reflection averaging (agreement on I = 4.2%)
Observations:	973 with I>3σ (I)
Parameters:	82
R:	0.094
wR:	0.098
Goodness-of-fit:	4.602
Maximum shift/error:	0.0009
Residual density	
maximum:	0.5 e/Å^3
minimum:	-0.5 e/Å^3

Table 2 Interatomic Bond Distances, Angles and Dihedral Angles for p-Bis(phenyl)tetrazine.

Atom	Distance(Å)
N(1)-C(2)	1.354(5)
N(1)-N(3')	1.316(5)
C(2)-C(4)	1.450(5)
C(2)-N(3)	1.346(5)
C(4)-C(5)	1.398(5)
C(4)-C(9)	1.402(5)
C(5)-C(6)	1.369(6)
C(5)-H(5)	1.040(5)
C(6)-C(7)	1.401(6)
C(6)-H(6)	1.040(6)
C(7)-C(8)	1.385(6)
C(7)-H(7)	1.040(7)
C(8)-C(9)	1.366(6)
C(8)-H(8)	1.040(6)
C(9)-H(9)	1.040(5)

Atoms	Angle(°)
C(2)-N(1)-N(3')	117.9(3)
N(1)-C(2)-C(4)	117.9(3)
N(1)-C(2)-N(3)	123.2(4)
C(4)-C(2)-N(3)	118.9(3)
N(1')-N(3)-C(2)	118.9(3)
C(2)-C(4)-C(5)	120.7(3)
C(2)-C(4)-C(9)	121.1(3)
C(5)-C(4)-C(9)	118.2(4)
C(4)-C(5)-C(6)	120.8(4)
C(4)-C(5)-H(5)	119.1(4)
C(6)-C(5)-H(5)	120.0(4)
C(5)-C(6)-C(7)	120.7(4)
C(5)-C(6)-H(6)	119.8(5)
C(7)-C(6)-H(6)	119.4(5)
C(6)-C(7)-C(8)	118.2(4)
C(6)-C(7)-H(7)	120.9(5)
C(8)-C(7)-H(7)	120.8(5)
C(7)-C(8)-C(9)	121.5(4)
C(7)-C(8)-H(8)	119.0(5)
C(9)-C(8)-H(8)	119.4(5)
C(4)-C(9)-C(8)	120.5(4)
C(4)-C(9)-H(9)	119.2(5)
C(8)-C(9)-H(9)	120.3(4)

Atoms	Dihedral Angle(°)
N(3')-N(1)-C(2)-C(4)	-179.6(3)
N(3')-N(1)-C(2)-N(3)	0.6(6)
C(2)-N(1)-N(3')-C(2')	-0.5(5)
N(1)-C(2)-C(4)-C(5)	-0.6(6)
N(1)-C(2)-C(4)-C(9)	-179.7(4)
N(3)-C(2)-C(4)-C(5)	179.3(4)
N(3)-C(2)-C(4)-C(9)	0.2(6)
C(2)-C(4)-C(5)-C(6)	-179.2(4)
C(2)-C(4)-C(5)-H(5)	-1.4(6)
C(9)-C(4)-C(5)-C(6)	-0.1(6)
C(9)-C(4)-C(5)-H(5)	177.8(4)
C(2)-C(4)-C(9)-C(8)	178.8(4)
C(2)-C(4)-C(9)-H(9)	1.1(7)
C(5)-C(4)-C(9)-C(8)	-0.3(6)
C(5)-C(4)-C(9)-H(9)	-178.0(5)
C(4)-C(5)-C(6)-C(7)	-0.2(7)
C(4)-C(5)-C(6)-H(6)	-177.2(5)
H(5)-C(5)-C(6)-C(7)	-178.1(5)
H(5)-C(5)-C(6)-H(6)	5.0(8)
C(5)-C(6)-C(7)-C(8)	0.9(7)
C(5)-C(6)-C(7)-H(7)	179.1(5)
H(6)-C(6)-C(7)-C(8)	177.8(5)
H(6)-C(6)-C(7)-H(7)	-4.0(8)
C(6)-C(7)-C(8)-C(9)	-1.3(7)
C(6)-C(7)-C(8)-H(8)	179.2(5)
H(7)-C(7)-C(8)-C(9)	-179.5(5)
H(7)-C(7)-C(8)-H(8)	1.0(8)
C(7)-C(8)-C(9)-C(4)	1.0(7)
C(7)-C(8)-C(9)-H(9)	178.7(5)
H(8)-C(8)-C(9)-C(4)	-179.5(5)
H(8)-C(8)-C(9)-H(9)	-1.8(8)

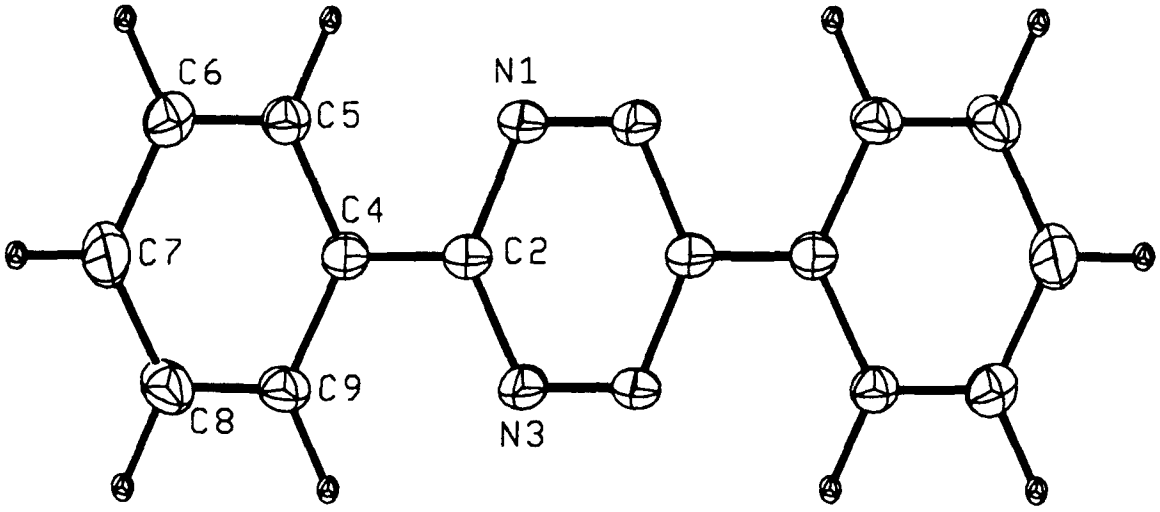


Figure 1. ORTEP drawing of p-Bis(phenyl)tetrazine. Non-hydrogen atoms are shown at 30% thermal probability. Hydrogen atoms are shown artificially small for clarity.

1717s

Eddy Current and Dielectric Spectroscopy Characterization of Metals and Ceramics - Application to NDE

FINAL REPORT

by

Prasad K. Kadaba
(Principal Investigator)

Xiaoming Lou
(Research Associate)

Contract No. : S - 210 - 11MG - 085
Contract Period : June 1, 1991 - December 31, 1991

Department of Electrical Engineering
UNIVERSITY OF KENTUCKY
LEXINGTON, KY 40506

ACKNOWLEDGEMENTS

The authors wish to acknowledge in particular the assistance of Drs Pramod K. Bhagat and Tom Moran in the Material Laboratory (WL/MLLP) of Wright-Patterson AFB during the progress of the project. They also wish to acknowledge the award of the grant by the Air Force Office of Scientific Research through the Universal Engergy Systems, Dayton, OH. The authors also wish to thank the University of Kentucky Center for Research and Manufacturing for the use of their automation equipment.

Abstract.

This report presents an eddy-current test set-up designed and constructed at the University of Kentucky during the seven month period of the project. We have demonstrated that second-layer cracks can be delineated to an accuracy of about 2 percent in materials of high conductivity using long transmitter pulse scheme. The pulse transmitter has been designed and built. In the theoretical analysis of electromagnetic NDE, the method of moments has been applied to calculate the variation of the probe impedance due to arbtrarily shaped defects in materials of high conductivity. A detailed FORTRAN computer program was developed for the use of this technique.

A small effort was addressed to study the microstructural state of alumina ceramics using the technique of Time-Domain Dielectric spectroscopy. Preliminary results seem promising for non-destructive evaluation (NDE) of ceramics. We plan to continue this effort in detail during the course of next year.

INTRODUCTION

Eddy-current testing is used in the field of nondestructive evalution to perform the inservice inspection of metal (conductor) products. The experienced operator can, in general, determine the presence of surface flaws, subsurface flaws particularly in nonmagnetic conductors, layer thickness and perform metal sorting (conductivity). In spite of its simplicity, currently, eddy current method is the most useful technique to detect the presence of indepth flaws especially in materials of high electrical conductivity such as metals. Another companion technique is the ultrasonic detection method, but eddy current method is non-contacting and the inspection can be carried out quickly with no contamination of the surface. Nondestructive techniques such as microwave scattering method although non-contacting is restricted to only surface characterization because of the skin-depth aspect of microwaves, especially in materials of high conductivity.

During the seven-month period of the minigrant, the following tasks have been addressed:

1. Development of an eddy-current pulse technique to explore second-layer cracks in materials of high conductivity.

2. Development of a computer program for the method of moments analysis to calculate the variation of probe impedance due to arbitrarily shaped defects in high conductivity materials.

3. Preliminary study of the microstructural state of sintered alumina by the technique of Time-Doman Dielectric Spectroscopy.

DETAILS OF THE INVESTIGATION

1. EDDY-CURRENT TEST SET-UP

In the literature, eddy-current test systems have been reported for different NDE tasks [1],[2],[3]. Speciallv in the area of NDE of a flat metal plate, many contributions have been published [4],···,[8]. Although some significant advances have been made in electromagnetic NDE, a pulse technique still needs to be developed in order to detect second-layer cracks or flaws in materials of high conductivity . A pulsed eddy-current test system for NDE of a conductor plate has been set up in our lab. The details are as follows.

In order to set up an eddy-current system, first of all, a probe needs to be designed for a special task. In designing a probe, of importance is its size i.e. its diameter and height, turns and the diameter of wires of coils, cored material and the arrangement of driving coil and receiving coil [9]. Using the above consideration, a probe has been designed for the eddy-current test system. some considerations in designing probes are introduced as follows.

At first, What parameters in electromagnetic NDE will be measured should be kept in the probe designer's mind, that is, whether reflection or transmission measurement is used , and whether impedances or output voltages / currents or output waveforms will be measured. In the case of reflection measurements, as shown in Fig. 1 (see page 15) , the probe generally consists of two coils, namely driving coil and receiving coil. However, in through-transmission measurements, as shown in Fig. 2 (see page 15) , these two coils are designed in two probes separately. In the reflection probe, the receiving coil usually has two sub-coils with opposite winding so that when the reflection probe is kept in the air, the output of the receiving coil can be nearly zero, i.e. self-zeroed. This is convenient to intercept measured data since the outputs of the receiving probe directly indicate the characterization of materials under test. However, this makes it complicated to design a reflection probe since a fine adjustment of the coil locations should be made.

In our reflection probe, there is only one coil, Because the impedance analyzer we are using has a high resolution and can be easily zeroed or calibrated by a computer through HP-IB. Our reflection probe can be used in both reflection and through-transmisison measurements.

In through-transmission measurements, the driving probe usually has the same size as the receiving probe. and is moved from one point to another as the receiving probe is moved. In this way, a mechnical scanner is required to have two headers to hold and move each probe (or in the other way, to move the heavy plate while two probes are fixed) . This machine was difficult to design and and build in the short period of the grant. An alternative way needs to be investigated . A technique to sove the above problem has been developed in this report. A fixed coil configuration, as shown in Fig. 3 (see page 16) , instead of a driving probe is used in this technique. In this configuration, coils are connected to each other in the way that the current directions of two adjacent coils are opposite in order to enlarge the magnetic fields between these two coils. On first thought, the fact that the magnetic fields produced by these coil currents maybe are non-uniform makes it difficult to detect defects in materials under test. Fortunately, however, this problem can be easily solved. The magnetic fields can

be not only theoretically calculated, but also calibrated with a standard metal plate. These coils can be designed using a thick wire which can carry a large current, and have different turns for some purposes. It should be also noted that these coils be placed far from other materials of high conductivity to avoid disturbing their magnetic fields.

The diameter of a wire used to make a coil depends on the root mean square current flowing in the coil. An empirical formula as follows is useful.

$$d \geq \frac{I_{rms}}{(\frac{5\pi}{4})}$$

where d is the diameter of wires in mm.

The number of turns of coils is limited by the size of probe and also depends on desired outputs of the probe. The coil of our probe has 400 turns.

Cored-material usually is air medium. If a ferrite rod is used for the core, the effects of magnetic saturation and hysteresis on the flux should be modeled [10].

For an impedance analyzer used in eddy-current test, of importance are its frequency band, resolution and accuracy etc. GR 1693 RLC DIGIBRIDGE is used in our eddy-current test system. It has a frequency band from 5 Hz to 200 kHz, a resolution of $4\frac{1}{2}$ and an accuracy of $\pm$ 0.01 %.

Pulse generator is one of the important units in a pulsed eddy-current test equipment. a commercial pulse generator with a suitable frequency band is available to us. Unfortunately, however, sine wave signals from GR 1693 RLC DIGIBRIDGE have a unstable phase drift since it is produced digitally. Unstable phase differences between the signals from GR 1693 and those from the commercial pulse generator make the readings of GR 1693 unstable. Thus a pulse circuit using the sine wave signals from GR 1693 as its inputs needs to be designed. Fig. 4 (see page 16) shows a block diagram of our pulse circuit which is listed in Appendix 1.

In fig. 4 , an follower is used to provide isolation between the GR 1693 and the pulse circuit assembly. Pulse generator has a variable resistance to change the duration of pulse signals . The magnitude of currents to the driving coil mainly depends on the power amplifier.

A block diagram of our eddy-current test system is shown in fig. 5 (see page 17) . A mechnical scanner (model ADM-1812/pc, manufactured by Creative Automation Company) is used and can be controlled to move in x, y or z direction by a computer through HP-IB. A program (see Appendix 2) has been developed to control the mechnical scanner to move in small steps in x or y direction. The output of the probe is stored in the computer and processed and then plotted. The experimental results will be given in section 3.

2. IMPEDANCE CALCULATION USING THE METHOD OF MOMENTS

In the area of eddy-current test, numerical methods of fast speed and good accuracy to analyze a forward problem, namely to predict probe response to a defect of known size, still need to be developed, although FEM and BEM are available in NDE. In this report, Moment Methods are applied in the analysis of a forward problem in eddy-current test.

Statement of the problem : Let us assume that there are some defects in a flat metal plate which is thick enough so that the field produced by the current in the probe coil can not transmit through it, as shown in fig. 6 (see page 18). As shown in fig. 7 (see page 18), the probe coil with the radius of 'a' is horizontally located in a height of 'h' in air. The axis of the probe is taken as the z-axis of the coordinate system and the surface of the metal plate as the x-y plane or the ρ–ϕ plane . Some arbitrarily shaped defects are centered at ρ_{oi}, ϕ_{oi}, Z_{oi} (i=1,2,.....,N) . The medium of the ith defect is characterized by the conductivity σ_{oi} , the permittivity ε_{oi} , and the permeability μ_{oi} . The metal plate is characterized by the parameters σ_1 , ε_1 and μ_1 .

The electromagnetic field produced by the coil current is disturbed by the defects because of the discontinuity on the interface between a defect and the metal plate. The total field is the superposition of the incident field existing when there are no defects and the scattered field due to defects. So the coil impedance also has two parts namely, a standard impedance presented by the incident field, and a disturbed impedance due to the scattered fields of the defects. The standard impedance can be calibrated or calculated using the method published in [11] . The disturbed impedance will be calculated in this report using the Method of Moments.

In order to obtain the disturbed impedance, it is necessary to calculate the scattered field by defects in the metal plate. According to the equivalence principle in electromagnetic theory [12] , the scattered field by defects can be equivalent to the field produced by the equivalent magnetic current sources on the interface between defects and the metal plate. Here, magnetic currents need to be impressed on the two sides of the interface in a way that both have the same magnitudes but the opposite directions in order to match the tangential component of the electric fields on the interface. The tangential component of the magnetic fields on the interface will be matched using the following equation.

$$(\vec{H}^s(\vec{M})+\vec{H}^{in})_{outside}=(\vec{H}^s(-\vec{M}))_{inside} \tag{1}$$

Then the fields both inside and outside defects are unique according to the uniqueness theorem in electromagnetic theory [12] . The above equation will be treated here using the point-matching technique with subsectional bases of Moment Methods in order to deal with problems with arbitrarily shaped defects. After the equivalent magnetic currents on the interfaces are obtained by solving equation (1), the scattered fields and the disturbed impedance can be calculated by the following equations.

$$\dot{E}^s = \nabla\times\dot{F} \tag{2}$$

$$Z^s = \int_{on\ coil} \dot{E}^s\cdot dl \tag{3}$$

where $\vec{F}$, an electric vector potential, will be treated in the following section, and in equation (3), the current flowing in the coil is taken to be a unity.

GREEN'S FUNCTIONS AND FIELD CONSIDERATIONS

The fields produced by the coil current has been given in reference [11], which can be written as follows in the metal plate

$$A_\phi=\frac{I\,a}{2\pi}\int_0^\infty J_1(\rho\xi)J_1(a\xi)\frac{\xi}{\lambda_1}T(\xi)e^{-\lambda_0 h+\lambda_1 z}d\xi \tag{4}$$

$$\vec{H}^{in}=\nabla\times\vec{A}_\phi \tag{5}$$

where

$$T(\xi)=\frac{2\mu_0\lambda_1}{\mu_0\lambda_1+\mu_1\lambda_0}$$

$$\lambda_1=\sqrt{\xi^2-k_1^2}$$

$$\lambda_0=\sqrt{\xi^2-k_0^2}$$

$$k_1=\omega\sqrt{\mu_1\varepsilon_1(1-j\frac{\sigma_1}{\omega\varepsilon_1})}$$

$$k_0=\omega\sqrt{\mu_0\varepsilon_0}$$

where I, the coil current, will be taken as unity in the following analysis and z is negative.

The field produced by a vertical electric diple in the metal can be obtained by reference [13] , which in air medium is given by :

$$A_z=\int_0^\infty J_0(\rho\xi)\frac{\xi}{\lambda_0}T'(\xi)e^{-\lambda_0 h+\lambda_1 z}d\xi \tag{6}$$

where

$$T'(\xi)=\frac{2\mu_1\lambda_0}{\mu_1\lambda_0+\mu_0\lambda_1}$$

The magnetic vector potential presented by a horizontal electric diple, in principle, has two components , one of which is in the direction of the diple, and the other in the z direction , because of the boundary conditions on the interface between the air and

the metal plate. In the problem treated by this report, z-component of the magnetic vector potential by the horizontal electric diple, however, is much smaller than the component in the direction of the diple since the metal plate has a high conductivity. For example, if the conductivity of the metal plate is the order of 10^6 the z-component of the magnetic vector potential will be about 10^{-3} of the component in the diple direction, and hence can be neglected. The component of the magnetic vector potential in the diple direction can be calculated by equation (6).

Using the duality of electromagnetic fields, we can get the electric vector potential from equation (6). Again, the z-component of the electric vector potential due to a horizontal magnetic diple is neglected for the same reason.

ANALYSIS USING THE METHOD OF MOMENTS

We can rewrite equation (1) as follows

$$\vec{H}^{s_l}_{outside}(\vec{M})+\vec{H}^{s_l}_{inside}(\vec{M})=-\vec{H}^{i\,s_l} \qquad (l=1,2,\cdots,N) \tag{7}$$

where

$$\vec{H}^{s_l}_{inside}(\vec{M}) = -\vec{H}^{s_l}_{inside}(-\vec{M}).$$

In order to treat a problem of arbitrarily shaped defects, we solve equation (7) for M using the point-matching method in conjunction with the subsectional bases.

Magnetic currents on the surface of the nth defect can be expanded by subsectional bases.

$$\vec{M}_n = \sum_i \vec{M}_{ni}\delta(s-s_{ni}) \qquad (n=1,2,\cdots,N) \tag{8}$$

where $\delta(s-s_{ni})$ is a delta function at s_{ni}

$$\vec{M}=\sum_n \vec{M}_n=\sum_n\sum_i \vec{M}_{ni}\delta(s-s_{ni}) \qquad (i=1,2,\cdots,k_n) \tag{9}$$

Substituting eq (9) in eq (7), we have :

$$\sum_n\sum_i \vec{H}^{s_l}_{outside}(\vec{M}_{ni}\delta(s-s_{ni}))+\sum_i \vec{H}^{s_l}_{inside}(\vec{M}_{li}\delta(s-s_{li}))=-\vec{H}^{i\,s_l} \tag{10}$$

In the point-matching method, the testing function is taken as a delta function $\delta(s-s_{lj})$. Taking inner product of $\delta(s-s_{lj})$ and each side of (10), we have

$$\sum_n\sum_i \vec{H}^{s_l}_{outside}(\vec{M}_{ni}\delta(s-s_{ni}))+\sum_i \vec{H}^{s_l}_{inside}(\vec{M}_{lj}\delta(s-s_{li}))=-\vec{H}^{i\,s_l} \tag{11}$$

where $\vec{H}^{i\,s_q}$ can be obtained from (5). $\vec{H}^{s_q}_{outside}(\vec{M}_{ni}\delta(s-s_{ni}))$ and $\vec{H}^{s_q}_{inside}(\vec{M}_{li}\delta(s-s_{li}))$ are obtained as follows

The electric vector potential $\vec{F}^{s_q}_{ni}$ produced by the magnetic current $\vec{M}_{ni}\delta(s-s_{ni})$ outside the defect can be writen as follows

$$\vec{F}^{s_q}_{z/t\,ni\ outside} = \frac{M_{z/t\,ni}}{4\pi}\left[\int_0^{\infty} J_0(\xi|\rho_{lj}-\rho_{ni}|)\frac{\xi}{\lambda_1}e^{-\lambda_1(z_q+z_{ni})}\Gamma_{z/t}(\xi)d\xi+\frac{e^{-jk_1|r_{ni}-r_{lj}|}}{|r_{ni}-r_{lj}|}\right] \tag{12}$$

where

$$\Gamma_z(\xi)=\frac{\mu_0\lambda_1-\mu_1\lambda_0}{\mu_0\lambda_1+\mu_1\lambda_0}$$

$$\Gamma_t(\xi)=\frac{\varepsilon_1\lambda_0-\varepsilon_0\lambda_1}{\varepsilon_1\lambda_0+\varepsilon_0\lambda_1}$$

The magnitic field outside the defect is :

$$\vec{H}^{s_q}_{z/t_{ni}\ outside}=-j\omega\varepsilon_1\vec{F}^{s_q}_{z/t_{ni}\ outside}+\frac{1}{j\omega\mu_1}\nabla(\nabla\cdot\vec{F}^{s_q}_{z/t_{ni}\ outside}) \tag{13}$$

The electric vector potential inside the defect is :

$$\vec{F}^{s_q}_{z/t_{ni}\ inside}=\frac{M_{z/t\,ni}}{4\pi}\left[\frac{e^{-jk_{ol}|r_{ni}-r_{lj}|}}{|r_{ni}-r_{lj}|}\right] \tag{14}$$

The magnetic field inside the defect is given by :

$$\vec{H}^{s_q}_{z/t_l\ inside}=-j\omega\varepsilon_{ol}\vec{F}^{s_q}_{z/t_l\ inside}+\frac{1}{j\omega\mu_{ol}}\nabla(\nabla\cdot\vec{F}^{s_q}_{z/t_l\ inside}) \tag{15}$$

where the subscript "t" means parellel to the interface between the air and the metal plate .

Deduced from (13) and (15), $H_{z/t}$ both outside and inside the defect can be writen as follows

$$H^{s_q}_{v\ outside}(M_{ni}\delta(s-s_{ni}))=[A_{x_{ni}}]^v_{lj}M_{x_{ni}}+[A_{y_{ni}}]^v_{lj}M_{y_{ni}}+[A_{z_{ni}}]^v_{lj}M_{z_{ni}} \tag{16}$$

where

$$[A_{x_{ni}}]^v_{lj}=\begin{bmatrix}A_{xx_{ni}}\\A_{yx_{ni}}\\A_{zx_{ni}}\end{bmatrix}^v_{lj}$$

$$[A_{y_n}]^v_{ij}=\begin{bmatrix} A_{xy_n} \\ A_{yy_n} \\ A_{zy_n} \end{bmatrix}^v_{ij}$$

$$[A_{z_n}]^v_{ij}=\begin{bmatrix} A_{xz_n} \\ A_{yz_n} \\ A_{zz_n} \end{bmatrix}^v_{ij}$$

Here v stards for x, y or z.

$$H^{i_v}_{v__}(M_{li}\delta(s-s_{li}))=[B_{x_n}]^v_{ij}M_{x_n}+[B_{y_n}]^v_{ij}M_{y_n}+[B_{z_n}]^v_{ij}M_{z_n} \tag{17}$$

Substituting the results (16) and (17) in (11), we have

$$\sum_n\sum_i\{[A_{x_n}]^v_{ij}M_{x_n}+[A_{y_n}]^v_{ij}M_{y_n}+[A_{z_n}]^v_{ij}M_{z_n}\}+\sum_i\{[B_{x_i}]^v_{ij}M_{x_i}+[B_{y_i}]^v_{ij}M_{y_i}+[B_{z_i}]^v_{ij}M_{z_i}\}=-H^{i_{s_v}}_v \tag{18}$$

Using matrixe notation, we have :

$$[A_x]^v_{ij}[M_x]+[A_y]^v_{ij}[M_y]+[A_z]^v_{ij}[M_z]+[B_{x_i}]^v_{ij}[M_{x_i}]+[B_{y_i}]^v_{ij}[M_{y_i}]+[B_{z_i}]^v_{ij}[M_{z_i}]=-H^{i_{s_v}}_v \tag{19}$$

where

$$[A_x]^v_{ij}=[[A_{x_1}]^v_{ij},[A_{x_2}]^v_{ij},\cdots,[A_{x_N}]^v_{ij}]$$

$$[A_{x_i}]^v=[A_{xi1},A_{xi2},\cdots,A_{xik_i}]^v$$

$$[M_x]=[[M_{x1}],[M_{x2}],\cdots,[M_{xN}]]^T$$

$$[M_{x_i}]=[M_{xi1},M_{xi2},\cdots,M_{xik_i}]^T$$

Letting $l=1,2,\ldots,N$, $j=1,2,\ldots,k_l$, we have

$$[A_x]^v[M_x]+[A_y]^v[M_y]+[A_z]^v[M_z]+[B_x]^v[M_x]+[B_y]^v[M_y]+[B_z]^v[M_z]=-[H^i_v] \tag{20}$$

where

$$[A_x]^v=[[A_x]^v_1,[A_x]^v_2,\cdots,[A_x]_N]^T$$

$$[A_x]^v_n=[[A_x]_{n1},[A_x]^v_{n2},\cdots,[A_x]_{nk_n}]^T$$

$$[B_z]^v=\begin{bmatrix}[B_{z1}]^v & & & \\ & [B_{z2}]^v & & \\ & & \ddots & \\ & & & [B_{z_n}]^v\end{bmatrix}$$

$$[B_{z_n}]^v=\begin{bmatrix}B_{z_{n1}} & B_{z_{n2}} & & B_{z1k_n} \\ \vdots & \vdots & & \vdots \\ B_{zk_n1} & B_{zk_n2} & & B_{zk_nk_n}\end{bmatrix}$$

$$[H_v^i]=[[H_v^i]_1,[H_v^i]_2,\cdots,[H_v^i]_N]^T$$

$$[H_v^i]_n=[H_{v1}^i,H_{v2}^i,\cdots,H_{vk_n}^i]^T$$

Eq (20) can be rewriten as follows :

$$[C_x]^v[M_x]+[C_y]^v[M_y]+[C_z]^v[M_z]=-[H_v^i] \tag{21}$$

v = x, y, z

where

$$[C_z]^v=[A_z]^v+[B_z]^v$$

By solving equation (21), we can obtain M_x, M_y and M_z .

IMPEDANCE CALCULATION

After the magnetic currents M_x, M_y and M_z are obtained, the impedance of the receiving probe due to them can be calculated thus :

Letting the current in the receiving coil be equal to unity, we can obtain the impedance by calculating the voltage induced in the receiving coil

$$Z=V=\int \vec{E}\cdot d\vec{l}=\int_0^{2\pi\rho} E_\phi dl \tag{22}$$

E_ϕ produced by M_x, M_y and M_z can be obtained as follows

As shown in fig. 8 (a) (see page 19), a magnetic current source with three components M_{xm},M_{ym}andM_{zm} is located at (x_m, y_m, z_m); Then the electric vector potentials F_{xm}, F_{ym} and F_{zm} can be obtained as follows

$$F_{v_{ni}}=\frac{M_{v_{ni}}}{4\pi}\int_0^{\infty} J_0(\xi\rho_{nig})\frac{\xi}{\lambda_0}T_v(\xi)e^{-h\lambda_0+\lambda_1 z_{ni}}d\xi \tag{23}$$

where

$$T_{x/y}(\xi)=\frac{2\varepsilon_1\lambda_0}{\varepsilon_1\lambda_0+\varepsilon_0\lambda_1}$$

$$T_z(\xi)=\frac{2\mu_1\lambda_0}{\mu_1\lambda_0+\mu_0\lambda_1}$$

$$\rho_{nig}=[(x_{ni}-a\cos\phi_g)^2+(y_{ni}-a\sin\phi_g)^2]^{\frac{1}{2}}$$

The electric fields can then be obtained from equation (23), which is

$$E_{v_{ni}}=\sum_v(\nabla\times\vec{F}_{v\,ni})\cdot\vec{v} \tag{24}$$

The receiving coil is assumed to be horizontally placed , so $E_{z\,ni}$ has no contribution to the voltage induced on the coil. $E_{x\,ni}$ and $E_{y\,ni}$ are of interest here. As shown in fig. 8(b) (see page 19),

$$E^g_{\phi_{ni}}=-E_{x_{ni}}\sin\phi_g+E_{y_{ni}}\cos\phi_g \tag{25}$$

The voltage (or impedance) due to the magnetic current source M_{ni} is

$$V^g_{ni}=E^g_{\phi_{ni}}\,\Delta l \tag{26}$$

where $\Delta l=\frac{2\pi a}{Q}$ (the coil is divided into Q sub-sections)

The total voltage (or impedance) is

$$Z=V=\sum_g\left(\sum_n\sum_i(V^g_{ni})\right) \tag{27}$$

PROGRAM

The flow chart of the FORTRAN program using the Method of Moments is shown in fig. 9 (see page 20). The detailed program is listed in Appendix 3.

CALCULATED RESULTS

We have used the above program to calculate the probe impedance variations with frequency from 60 Hz to 200 KHz when the probe is centered over two holes. Both holes have a circular cross section and a depth of 7.86 mm; their diameters are 6.35 mm and 3.81 mm, respectively. A comparison between the calculated results and the experimental results is shown in fig. 10 (see page 21). It is noted that a good agreement is presented between these two results.

3. EXPERIMENTAL INVESTIGATION

In this subsection, experimental studies of NDE of an aluminum plate have been reported. The experimental set-up, receiving probe and pulse current waveform to driving coils are shown in fig. 11 (see page 22). Three machined holes #1, #2, #3 in the plate are shown in fig. 12 (see page 23). In fig. 12, the sizes of the plate and three holes are also presented. The thickness of an individual aluminum foil used in the experiment is 0.2 mm. The operating frequency is 10 KHz.

In figs. 13 through 16, the eddy-current test results (the magnitudes of probe impedances) have been shown. In the experiment, after the mechnical scanner moves the probe by steps of 0.1 inch in the horizontal direction, and 0.5 inch in the vertical direction, the impedance readings of GR 1693 RLC DIGIBRIDGE is stored in the computer. In accordance with this, in each figure (14 through 16), there are 30 points in the horizontal direction and 5 rows in the vetical direction. As shown in the figures, impedance magnitude exhibits a peak when the probe is moved across the hole (defect). In figure 13, the hole is not covered by a layer of aluminum foil; so the impedance peaks are big in intensity under condition of both low and high resolution, while as in figures 14 and 15, the peaks are reduced in intensity with three layers of aluminum foil covering both hole #1 and #2. In figure 16 when 4 layers of aluminum foil are over hole #2, the impedance peaks are even reduced further. Fig. 17 shows preliminary results of detection of a second-layer crack using the current pulse technique.

The results presented in this report are with a current pulse of 0.5 amp. Longer current pulses are, of course, needed for detection of defects or flaws to greater depth in high conductivity materials. Current pulses of 50 amp are under development and the results will be forth coming early next year.

4. MICROSTRUCTURAL STATE OF ALUMINA CERAMICS BY THE TECHNIQUE OF TIME-DOMAIN DIELECTRIC SPECTROSCOPY (TDDS)

Approximately a month was spent in this aspect of the grant. Sintered alumina samples were measured in the frequency range 1 to 10 KHz at a temperature of $60^{\circ}c$. The experimentally measured parameter was the capacitance measured in a guarded-disk geometry by TDDS . The samples showed a fairly large change in ε^{1}, the real part of the permitivity with densification. Broad dispersion regions have been observed which are indicative of multiple dielectric relaxation processes. We plan to continue this study during next year at different temperatures and also at frequencies lower than 1 kHz both for pure alumina and glass-containing samples. Detailed interpretation of the results will lead to a characterization of different phases and interfaces present in the samples.

CONCLUSIONS

During the tenure of the mini-grant , a rather sophisticated pulse eddy-current test set-up has been developed. It has been demonstrated that second-layer cracks can be delineated to an accuracy of about 2 percent in aluminum metal plates. A pulse transmitter with 50 amp pulse current is almost complete and it is planned to continue this study next-year for detection of deeper flaws in non-magnetic as well as magnetic materials. The eddy-current test problem has been analyzed using the method of moments and a detailed FORTRAN program specific to the present investigation has been developed in-house.

Time-Domain Dielectric Spectroscopy (TDDS) has been used in a brief study of the microstructural state of alumina ceramics. Preliminary results are promising and the study will be continued next year extending to glass-containing samples.

REFERENCES

[1]. C.V.Dodd, L.M.Whitaker and W.E.Deeds, " An Accurate Laboratory Test System Using Commercial Equipment for Eddy-Current Measurements ", Materials Evaluation , Nov. 1988, PP. 1569-1575.

[2]. G.Wittig and H.M.Thomas, " Design of a Pulsed Eddy-Current Test Equipment with Digital Signal Analysis ", Eddy-Current Characterization of Materials and Structures, Birnbaum/Free, Ed., ASTM Publication 722, Philadelphia, 1981, pp. 375-387.

[3]. Jeff C. Treece, Thomas M. Roberts, Denis J.Radeeki and Steven D. Schunk, " Detecting Micro-Structure and flaws in Composites Using Eddy-Current Instrumentation " , 1988, pp.1519-1526.

[4]. C.V.Dodd and W.E.Deeds , " Absolute Eddy-Current Measurement of Electrical Conductivity " , Review of Progress in Quantitative Nondestructive Evaluation, Vol. 1, 1982, pp 387-394.

[5]. W.Lord, et al. , " A Finite Element Study of the Remote Field Eddy-Current Phonomenon " , IEEE Transactions on Magnetics, Vol. 24, Jan. 1988, PP. 435-438.

[6]. D.L.Waidelich , " Pulsed Eddy-Current Testing of Steel Sheets " , Eddy-Current Characterization of Materials and Structures , ASTM STP 722, Birnbaum/Free, 1981, pp.367-373.

[7]. S.K.Burke and L.R.F.Rose, " Interaction of Induced Currents with Cracks in Thin Plates ", Proc. R. Soc. Lond. A418, 1988, PP.229-246.

[8]. Sather Allen, " Investigation into the Depth of Pulsed Eddy-Current Penetration ", Eddy-Current Characterization of Materials and Structures , Birnbaum/Free, Ed., ASTM STP 722, 1981, PP.362

[9]. H.L.Libby, Introduction to Electromagnetic Nondestuctive Test Methods, Richland, Washington, Wiley-Interscience Publishers, 1971.

[10]. S.Herman and R.S.Prodan, " A Macroscopic Model of Eddy-Currents " , Eddy-Current Characterizatin of Material and Structures, Ed., ASTM STP 722, 1981, pp.86-93.

[11]. Cheng, D. " The Reflected Impedance of a Circular Coil in the Proximity of a Semi-Infinite Medium " , Ph.D Dissertation, University of Missouri, Jan. 1964.

[12]. R.F.Harrington , Time-Harmonic Electromagnetic Fields, McGRAWL HILL BOOK COMPANY, New York, 1961, pp.95-132.

[13]. Volkert W. Hansen, Numerical Solution of Antennas in Layered Media, John Wiley & Sons Inc., New York, 1989, pp. 13-40.

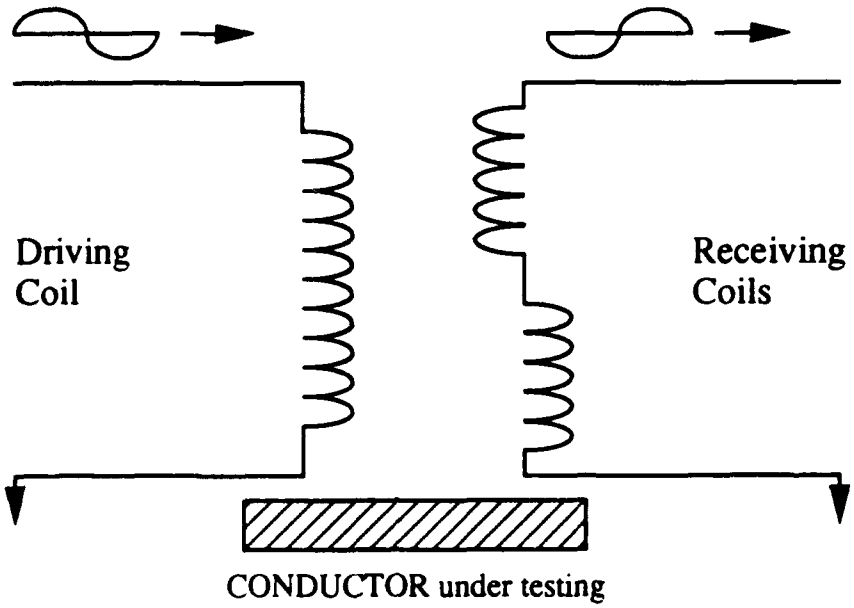


Fig. 1 Electrical connections for reflection measurement

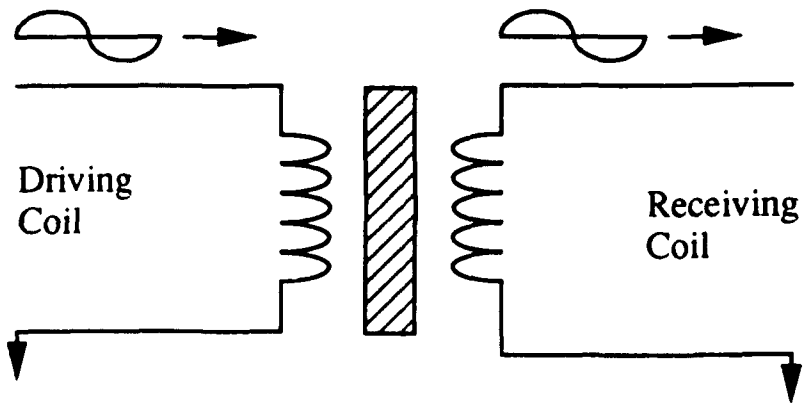


Fig. 2 Electrical connections for Through-transmission measurements

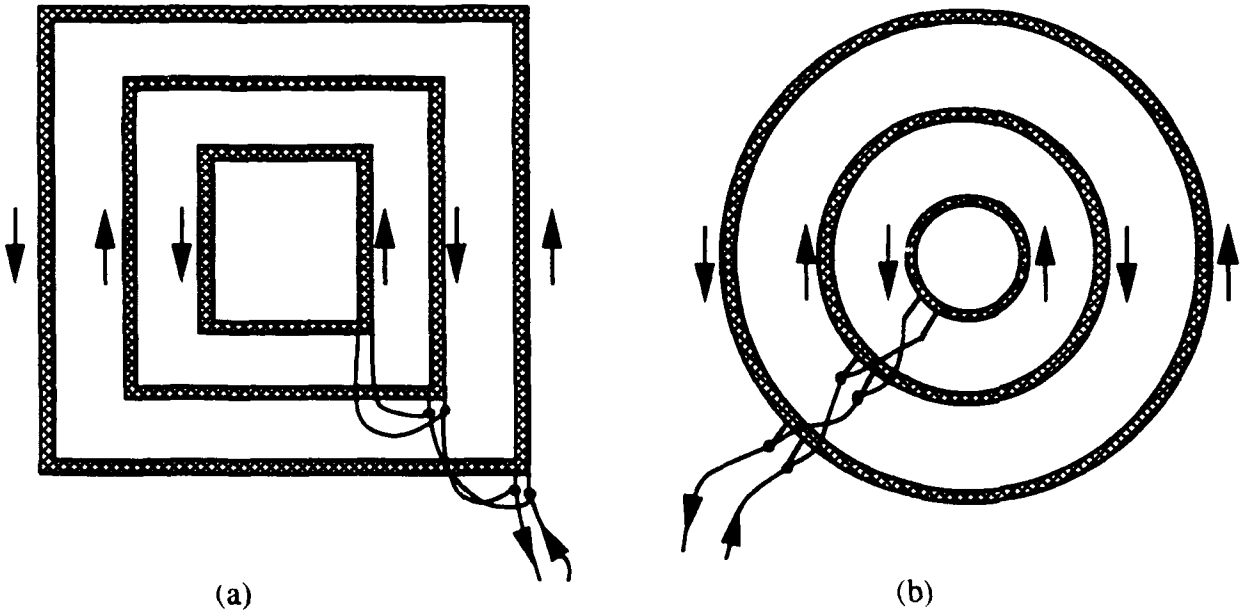


Fig. 3 Configurations for driving coils

(a) Rectangular
(b) Circular

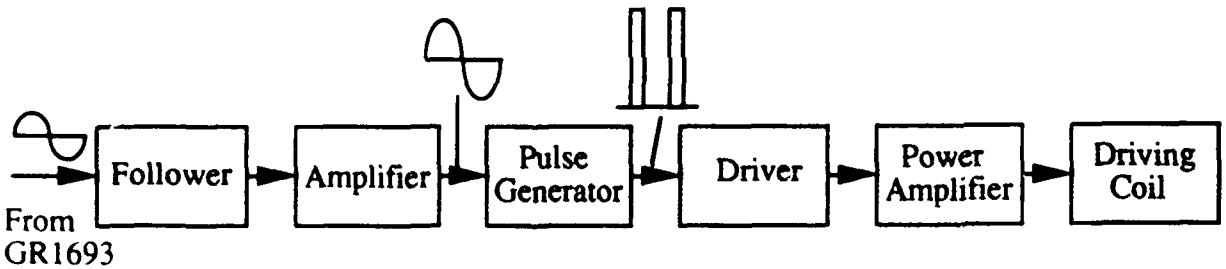


Fig. 4 Block Diagram for pulse circuit

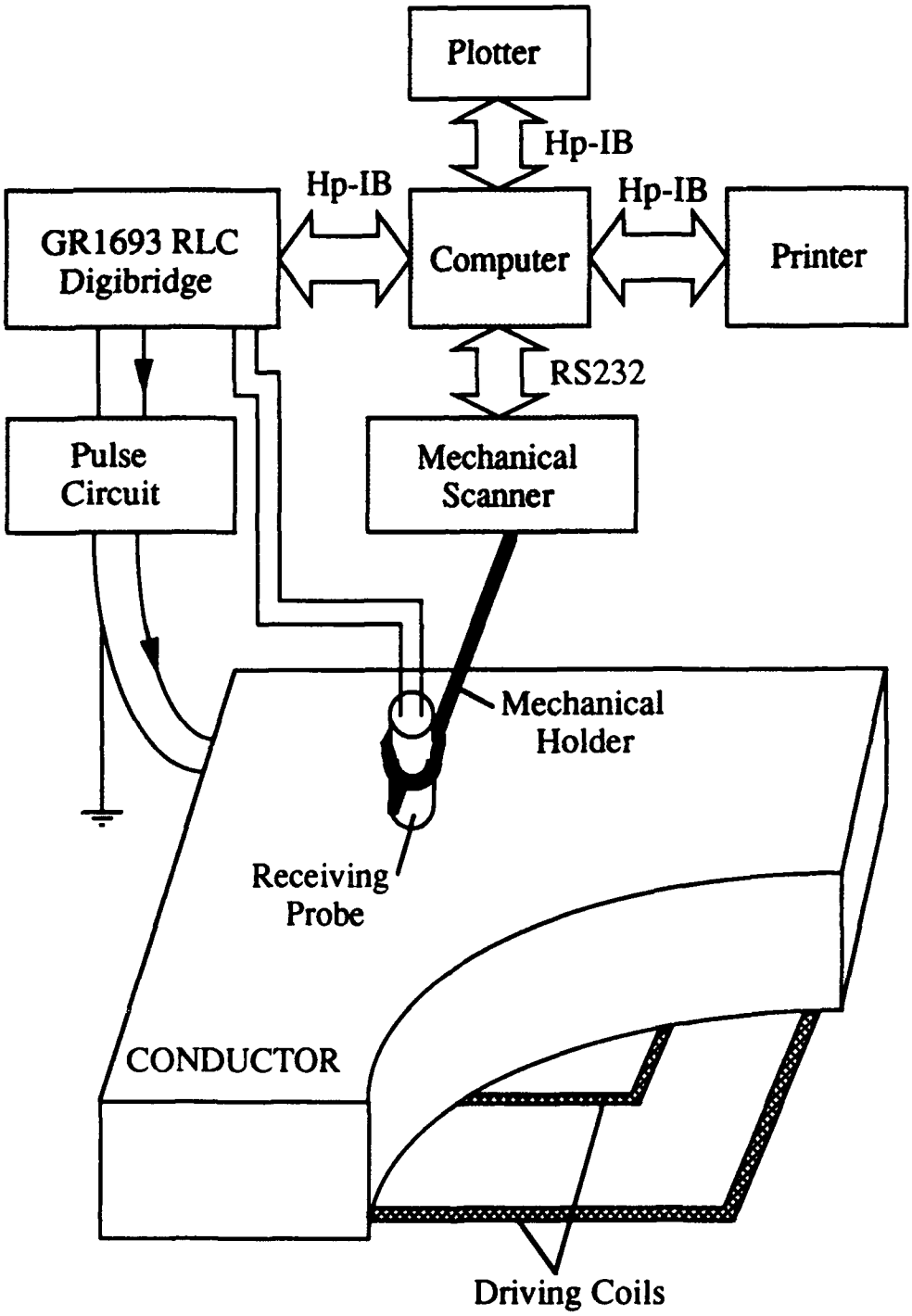


Fig. 5 Block diagram of eddy-current test system

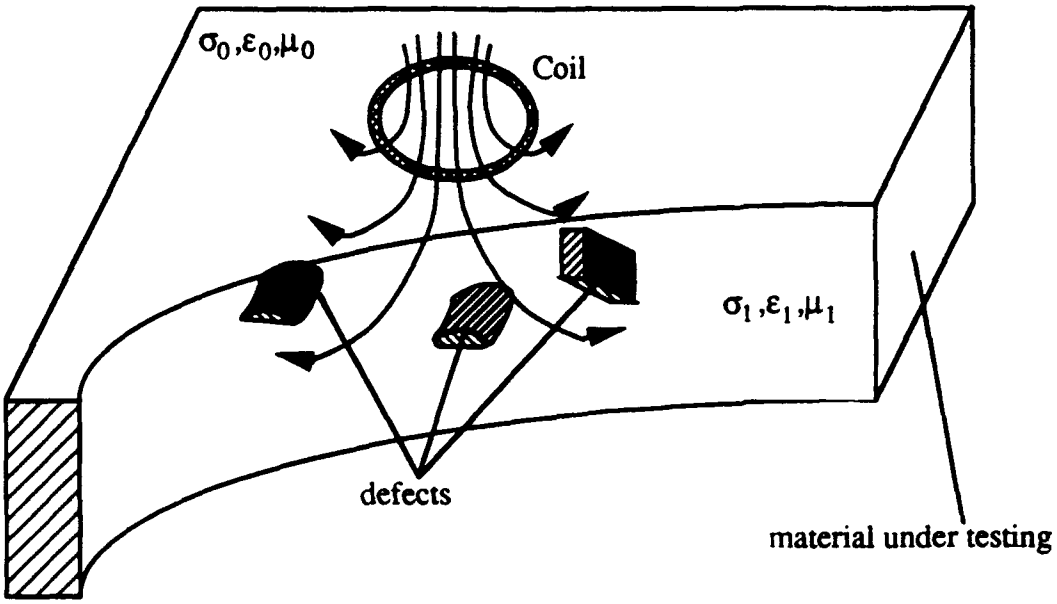


Fig. 6 Configurations for a practical problem

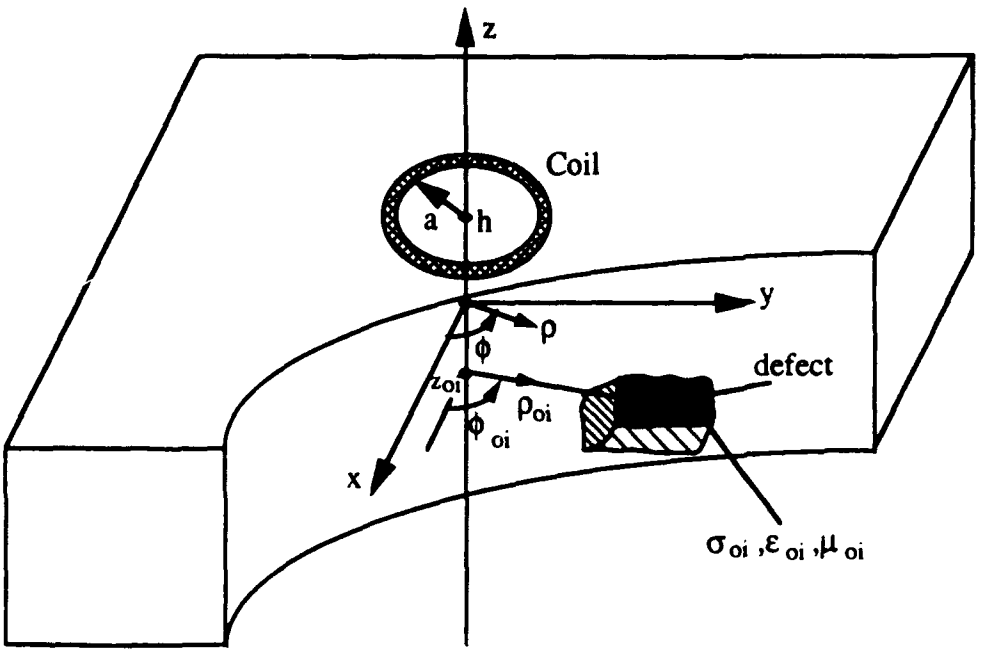
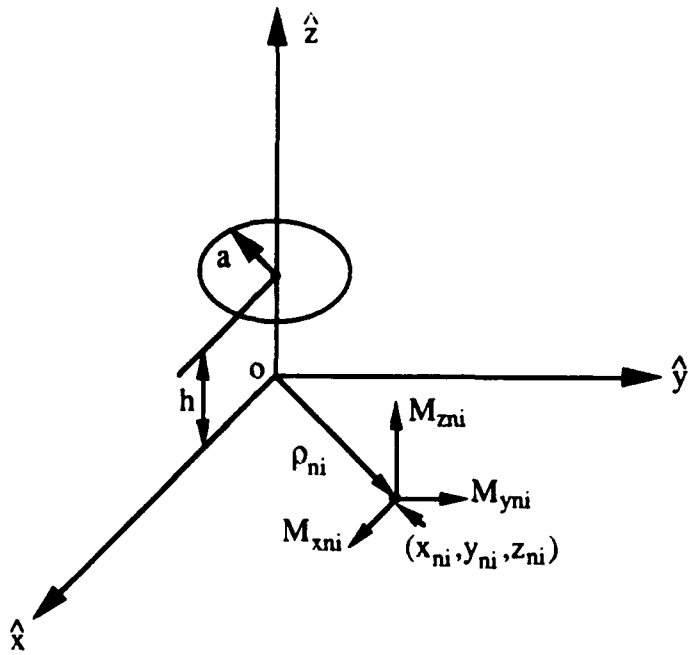
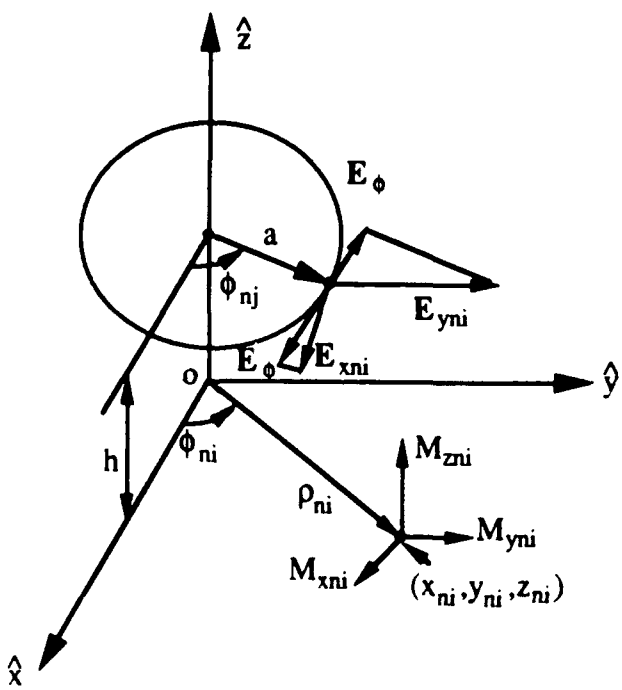


Fig. 7 The problem to be analyzed



(a)



(b)

Fig. 8 (a) Magnetic currents at S_{ni}
(b) Electric fields on the coil
produced by magnetic currents at S_{ni}

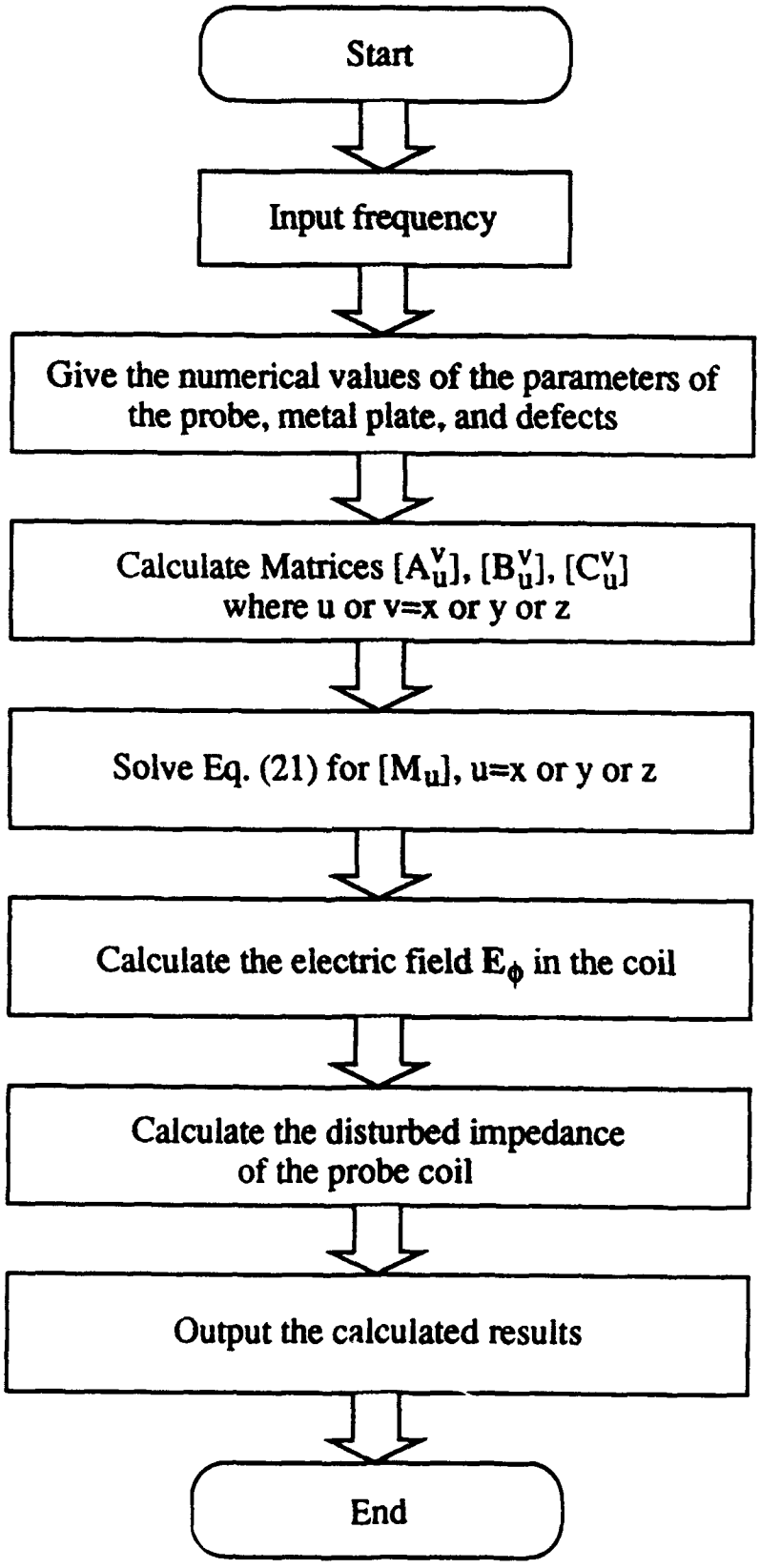


Fig. 9 FORTRAN program flow chart for calculating the coil impedance using the Method of Moments

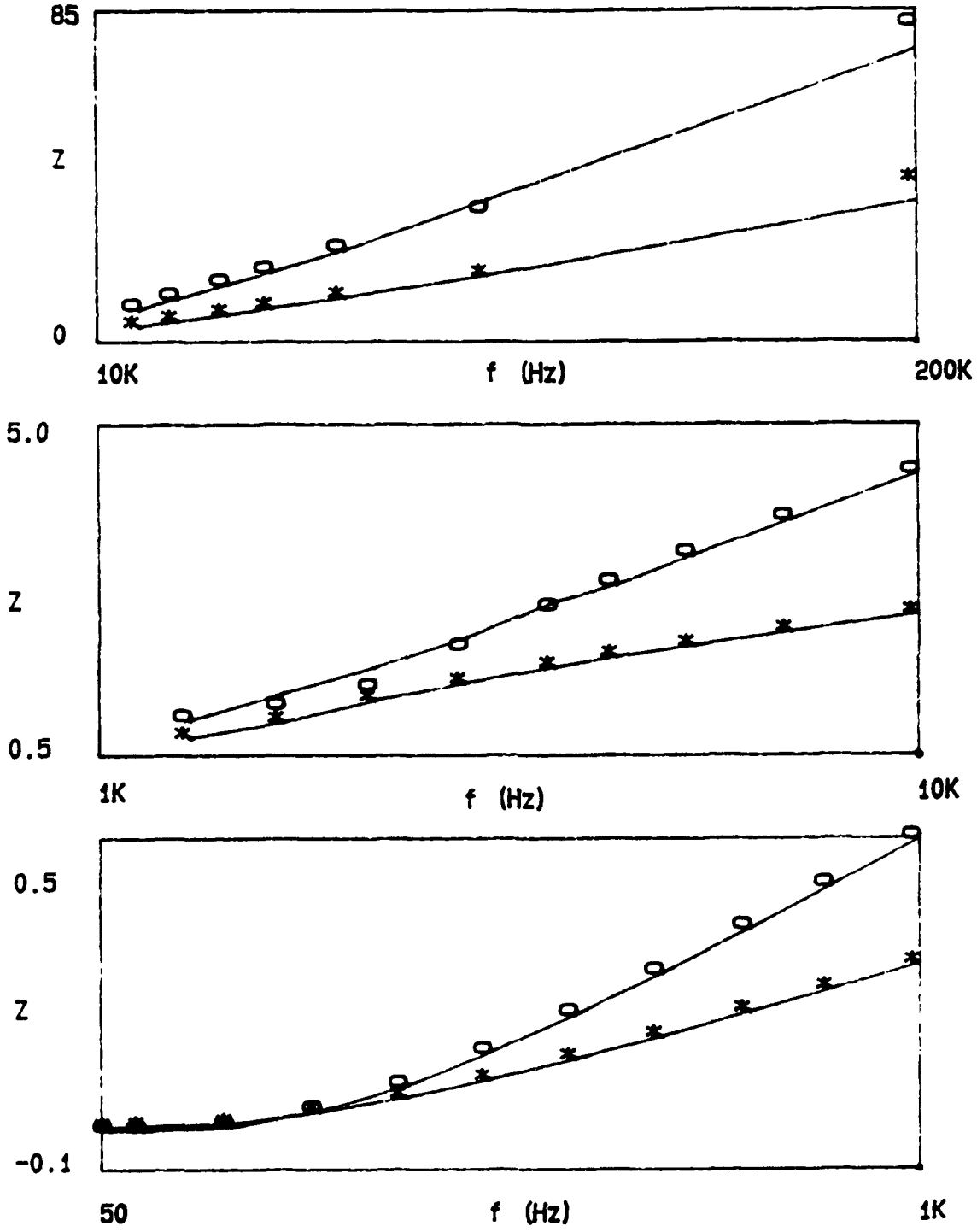


Fig. 10 Plot of the magnitude of impedance Z vs frequency in Hz

'O' experimental data with hole diameter, d=6.35 mm

'*' experimental data with hole diameter, d=3.81 mm

Solid lines for theoretical results in both above cases

Fig. 11(c) the current pulse waveform

Fig. 11 . (a) the eddy-current test set-up
(b) the detector probe
(c) the current pulse waveform

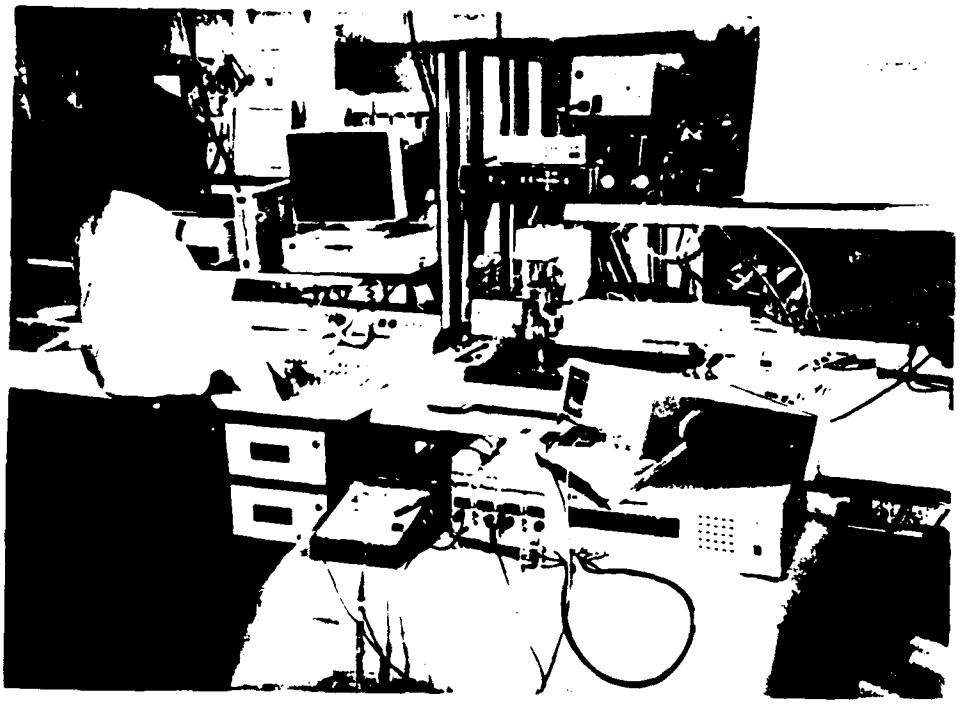


Fig. 11(a) the eddy-current test set-up

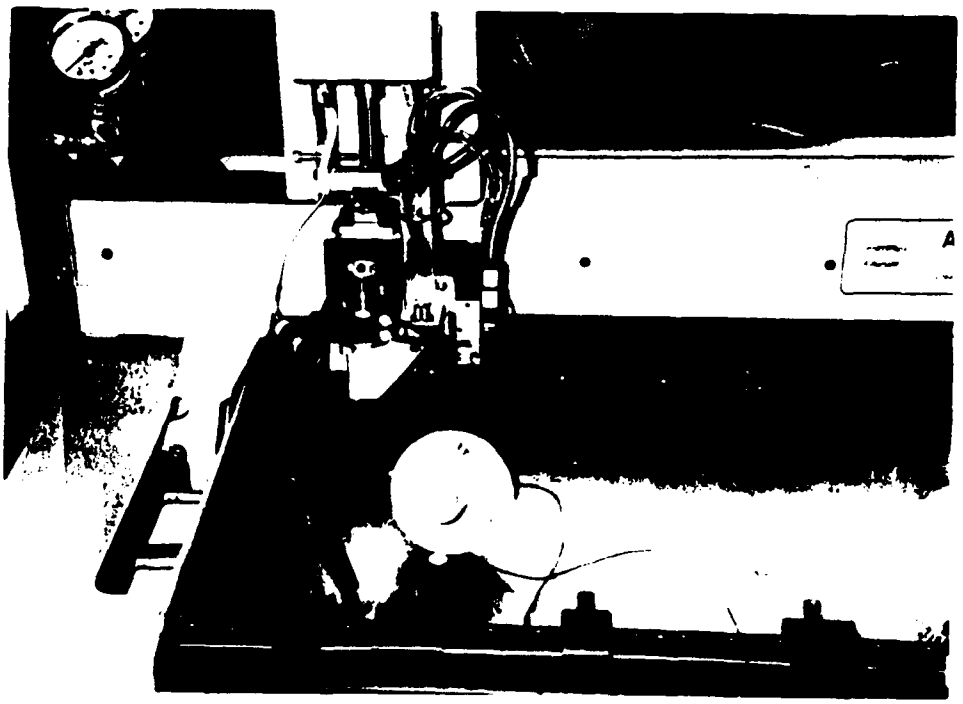


Fig. 11(b) the detector probe

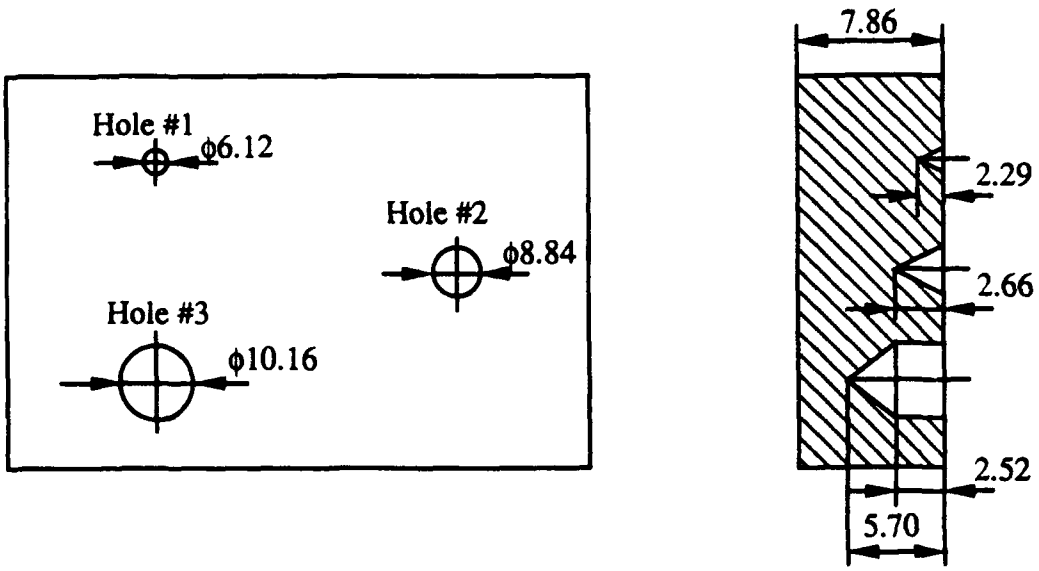


Fig. 12 Sizes and shapes of holes (units are in milimeters)

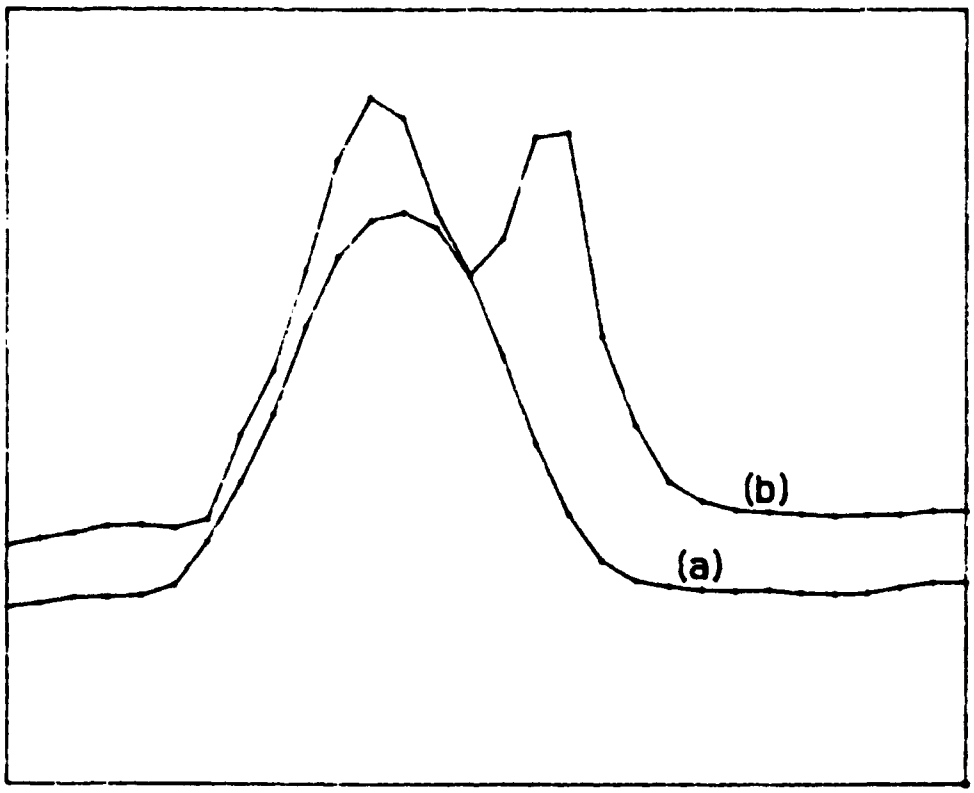


Fig. 13 No aluminum foil is placed on the hole #1

(a) under low resolution

(b) under high resolution

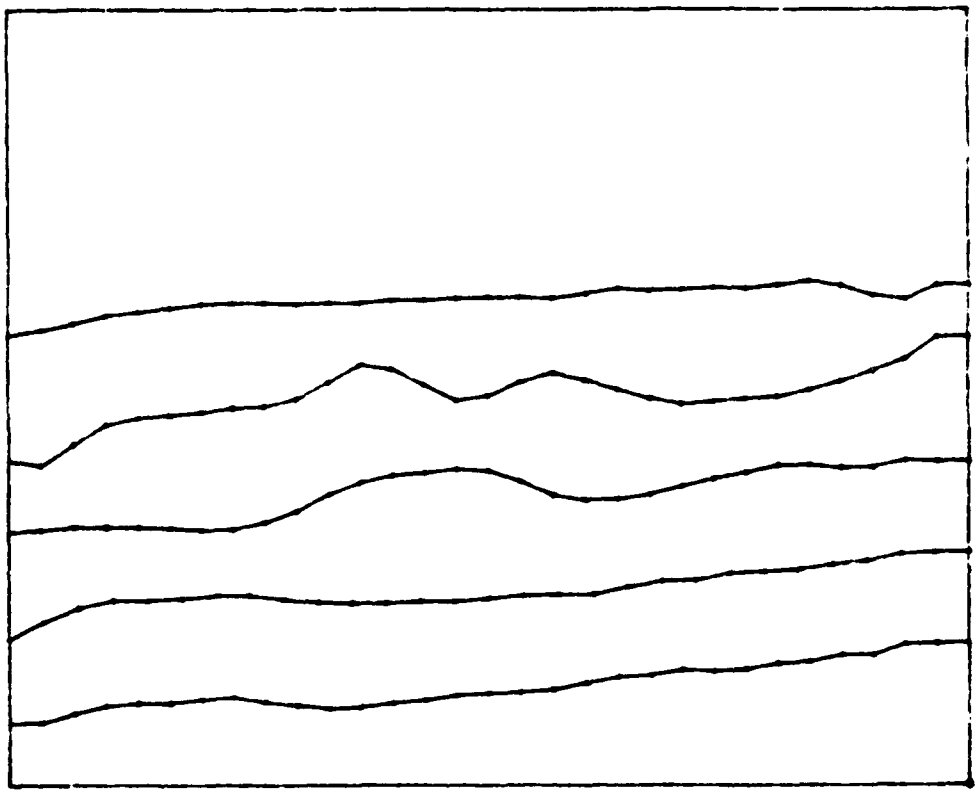


Fig. 14 Three layers of aluminum foils,

each of thickness 0.2 mm,

are placed on the hole #1.

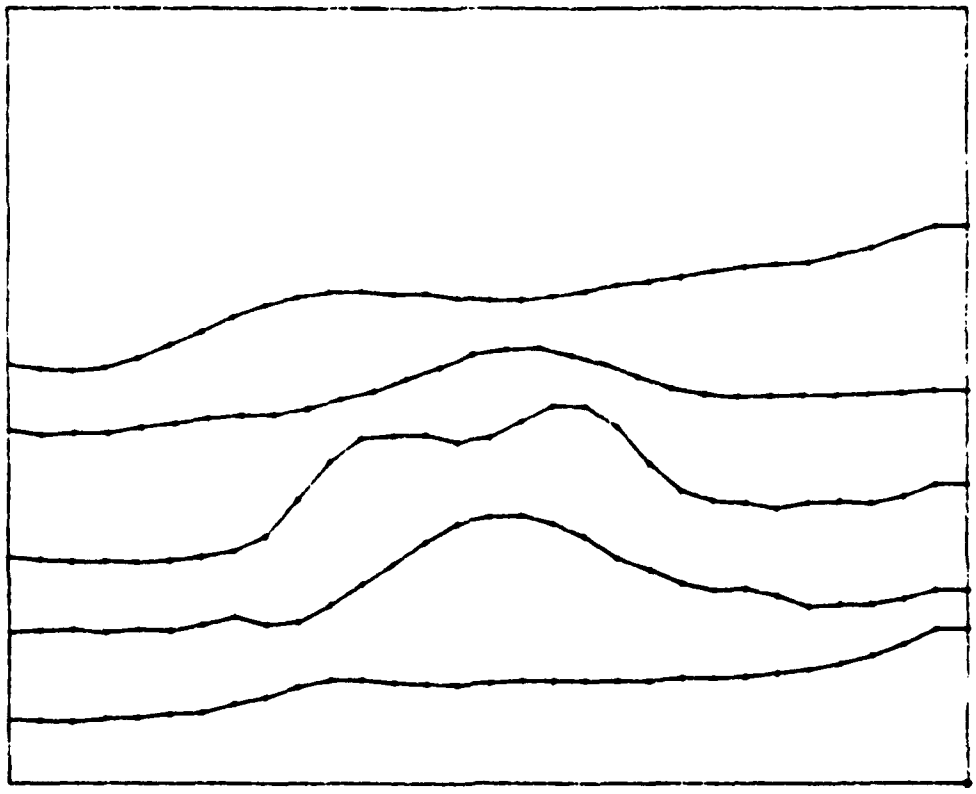


Fig. 15 Three layers of aluminum foils,

each of thickness 0.2 mm,

are placed on the hole #2.

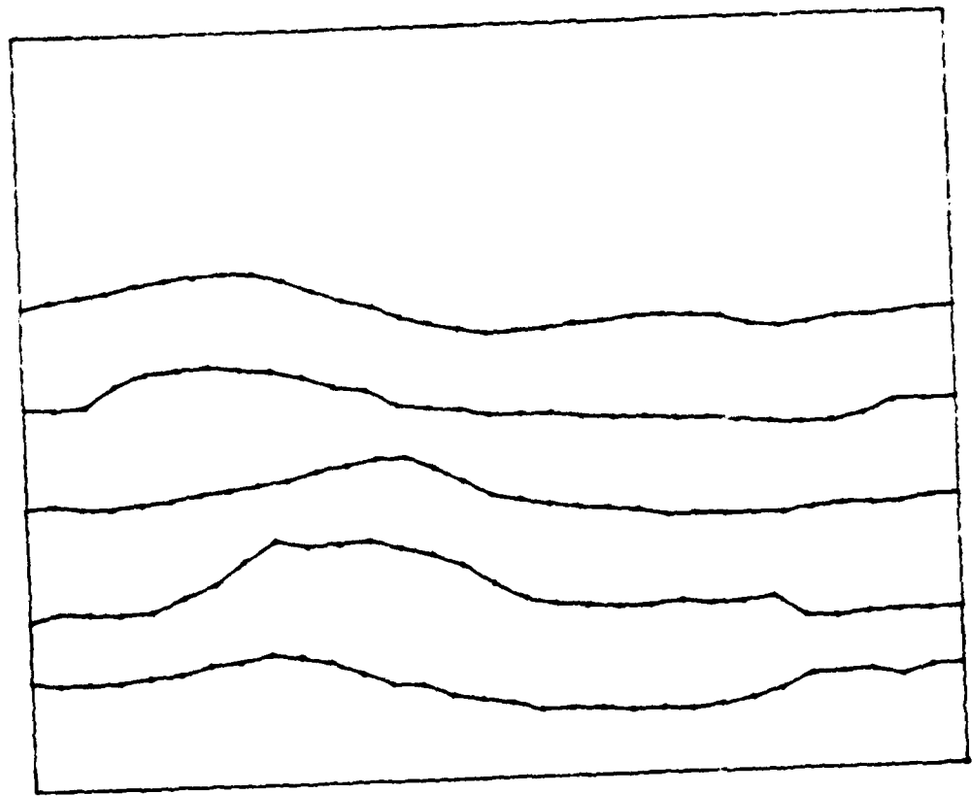


Fig. 16 Four layers of aluminum foils ,

each of thickness 0.2mm ,

are placed on the hole #2 .

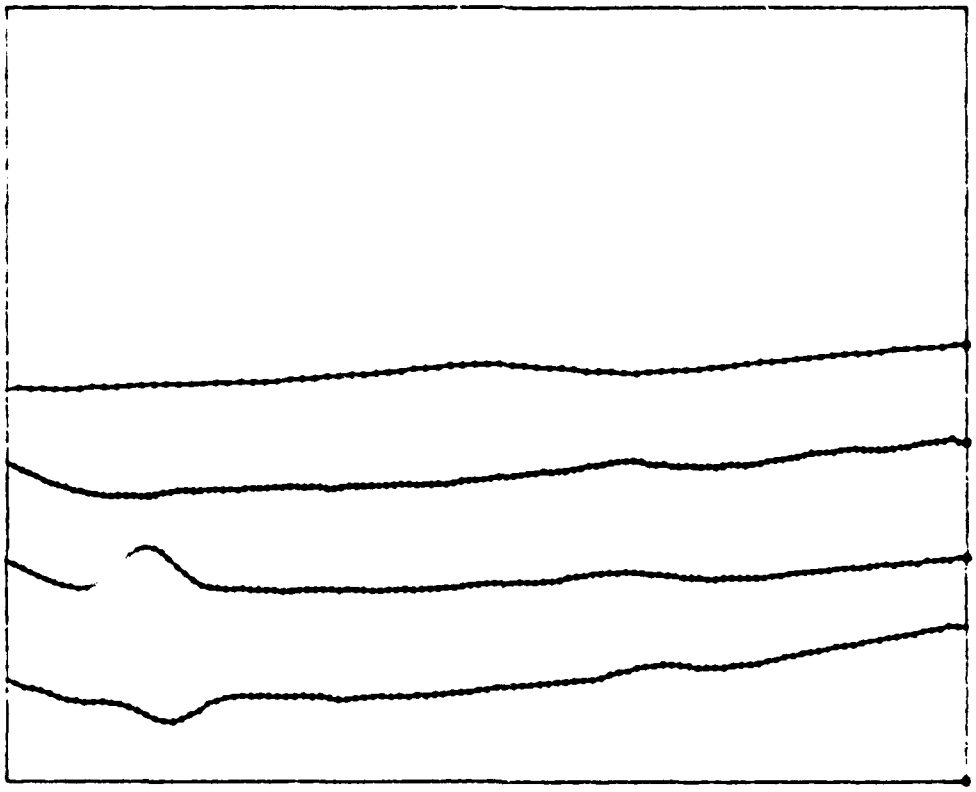
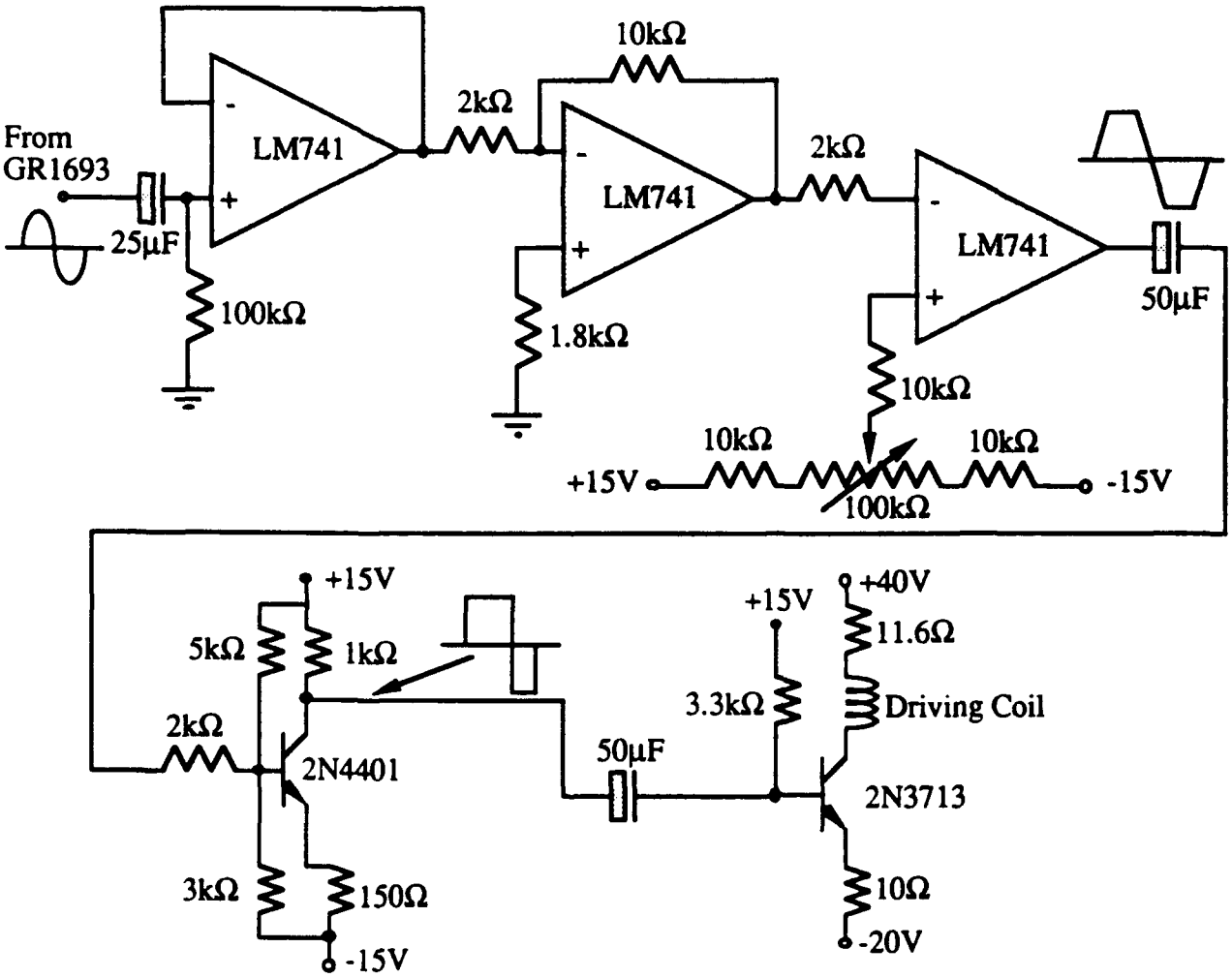


Fig. 17 Detection of a second layer crack using the current pulse technique

Appendix 1

Schematic of the Current Pulse Generator



Appendix 2

The Program Used to Control the Mechnical Scanner

Appendix 3

The FORTRAN Program
Used to Calculate the Receiving Coil Impedance Variations
Due to Defects or Flaws in Metal Plates
Using the Method of Moments

```
C ----------------  RETURN  WITH THE VALUE OF H(RUO)-----------------  EDD00010
C ---------------- RETURN WITH THE VALUE OF H(Z) --------------------   EDD00020
      PARAMETER (KKK=152)                                               EDD00030
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,XX,WW,NCOEFEDD00040
      COMMON BB,HHZ,RUO0,DD                                             EDD00050
      COMMON /WORKSP/  RWKSP                                            EDD00060
      COMPLEX HFIA(KKK),HZ(KKK),AZZ(KKK,KKK),AFF(KKK,KKK)               EDD00070
      COMPLEX BFF(KKK,KKK),BZZ(KKK,KKK),Y12,Z12,C(KKK,KKK),D(KKK,KKK)   EDD00080
      COMPLEX E(KKK,KKK),EFIA,SEFIA,COST1,COST2,RESULT1,RESULT2,A1      EDD00090
      REAL RWKSP(13944)                                                 EDD00100
      REAL K0,K1                                                        EDD00110
      CALL IWKIN(13944)                                                 EDD00120
      NCOEF=21                                                          EDD00130
      NNF=11                                                            EDD00140
      NNZ=3                                                             EDD00150
      CALL PARAM (NNF,NNZ,HFIA,HZ,AZZ,AFF,BZZ,BFF)                      EDD00160
      DO 051 I=0,NNZ                                                    EDD00170
      DO 052 J=1,NNF                                                    EDD00180
      IJ=I*NNF+J                                                        EDD00190
C     WRITE (6,*)HFIA(IJ),'------------HFIA'                           EDD00200
C     WRITE (6,*)HZ(IJ),'------------HZ',IJ                            EDD00210
052   CONTINUE                                                          EDD00220
051   CONTINUE                                                          EDD00230
      DO 061 I=0,NNZ                                                    EDD00240
      DO 062 J=1,NNF                                                    EDD00250
      IJ=I*NNF+J                                                        EDD00260
      DO 063 II=0,NNZ                                                   EDD00270
      DO 064 JJ=1,NNF                                                   EDD00280
      IIJ=II*NNF+JJ                                                     EDD00290
C     WRITE (6,*)BZZ(IJ,IIJ),'----------------- BZZ',IJ,IIJ            EDD00300
C     WRITE (6,*)BFF(IJ,IIJ),'----------------- BFF',IJ,IIJ            EDD00310
064   CONTINUE                                                          EDD00320
063   CONTINUE                                                          EDD00330
062   CONTINUE                                                          EDD00340
061   CONTINUE                                                          EDD00350
      LDA=NNF*(NNZ+1)                                                   EDD00360
C  ---------------------------------------------------------------------  EDD00370
C     LDA=2                                                             EDD00380
C     BFF(1,1)=CMPLX(1.,1.)                                             EDD00390
C     BFF(1,2)=CMPLX(1.,-1.)                                            EDD00400
C     BFF(2,1)=CMPLX(0.,-1.)                                            EDD00410
C     BFF(2,2)=CMPLX(0.,1.)                                             EDD00420
C     BZZ(1,1)=BFF(1,1)                                                 EDD00430
C     BZZ(1,2)=BFF(1,2)                                                 EDD00440
C     BZZ(2,1)=BFF(2,1)                                                 EDD00450
C     BZZ(2,2)=BFF(2,2)                                                 EDD00460
      CALL CGEMUL (BFF,KKK,'C',BFF,KKK,'N',C,KKK,LDA,LDA,LDA)           EDD00470
      CALL LINCG (LDA,C,KKK,C,KKK)                                      EDD00480
      DO 069 KJ=1,LDA                                                   EDD00490
      DO 068 KL=1,LDA                                                   EDD00500
C     WRITE(6,*)C(KJ,KL),KJ,KL                                          EDD00510
068   CONTINUE                                                          EDD00520
069   CONTINUE                                                          EDD00530
      CALL CGEMUL (C,KKK,'N',BFF,KKK,'C',E,KKK,LDA,LDA,LDA)             EDD00540
      CALL CGEMM ('N','N',LDA,LDA,LDA,(-1,0),E,KKK,BZZ,KKK,(0,0),C,KKK) EDD00550
      A1=CMPLX(-NNF*NNZ/BB/HHZ*2.,0.)                                   EDD00560
      CALL CGEMM ('N','N',LDA,1,LDA,A1,E,KKK,HFIA,KKK,(0,0),D,KKK)      EDD00570
      CALL CGEMM ('N','N',LDA,1,LDA,(-1,0),AFF,KKK,D,KKK,A1,HZ,KKK)     EDD00580
      CALL CGEMM ('N','N',LDA,LDA,LDA,(1,0),AFF,KKK,C,KKK,(1,0),AZZ,KKK)EDD00590
      CALL LSACG (LDA,AZZ,KKK,HZ,1,HFIA)                                EDD00600
```

```
      CALL CGEMM ('N','N',LDA,1,LDA,(1,0),C,KKK,HFIA,KKK,(1,0),D,KKK)
C ------------ MZ---HFIA,   MFIA----D ----------------------------
      DO 025 II=1,LDA
C     WRITE(6,*)HFIA(II),'--------------- MZ !'
025   CONTINUE
      DO 026 II=1,LDA
C     WRITE(6,*)D(II,1),'--------------------MFIA ! '
026   CONTINUE
C     END
C ------------ CALCULATE VOLTAGE ----------------------------
      SEFIA=CMPLX(0.,0.)
      DO 081 I3=1,NNF
      EFIA=CMPLX(0.,0.)
      FIA0=2*3.14159256*I3/NNF
      DO 071 I1=0,NNZ
      Z=DD+HHZ*I1/NNZ
      Z=-Z
      DO 072 I2=1,NNF
      JJ=I1*NNF+I2
      COST1=D(JJ,1)/4./3.14159256
      COST2=HFIA(JJ)/4./3.14159256
      FIA=2*3.14159256*I2/NNF
      XXI=BB*COS(FIA)
      YYI=BB*SIN(FIA)
      RUO2=RUO0**2+BB**2+2.*RUO0*BB*COS(FIA)
      RUO=RUO2+AA**2+2*AA*SQRT(RUO2)*COS(FIA0)
C ---------------- ERUO  , EFIA -------------------------
      RUO=SQRT(RUO)
      RUO2=SQRT(RUO2)
      CAA=AA
      CRUO=RUO
      CONDT=1.
      Z0=K0**2
      CALL HRUOZ (HR1,HX1,HR2,HX2,CONDT)
      RUO=CRUO
      AA=CAA
      RESULT1=CMPLX(HR1,HX1)
      RESULT2=CMPLX(HR2,HX2)
C     WRITE (6,*) RESULT1,RESULT2,'--------------- R1,R2',JJ
      EFIA=EFIA-RESULT1*COST1*(RUO2/RUO)*SIN(FIA0)
      EFIA=+EFIA+RESULT2*COST2*(AA+(RUO2)*COS(FIA0))/(RUO)
072   CONTINUE
071   CONTINUE
      SEFIA=SEFIA+EFIA*2*3.14159256*AA/NNF
081   CONTINUE
      WRITE (6,*)SEFIA,'--------------- IMPEDANCE'
      END
C ------------------ CALCULATE THE VALUE OF FIELD ----------------
      SUBROUTINE HRUOZ (HR1,HX1,HR2,HX2,CONDT)
      INTEGER  INTERV,NOUT,IRULE
      REAL  A,ABS,B,ALOG,ATAN,BOUND,ERRABS,ERREST,ERROR,ERRREL,EXACT
      REAL RESULT,CONST,K1,K0
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEF
      COMMON BB,HHZ,RUO0,DD
      INTRINSIC ABS,ALOG,SQRT,COS,SIN
      EXTERNAL F,G,DQDAG,UMACH,CONST,QDAG,FN1,FN2,FM1,FM2,FP,FP2,FA,FB
      EXTERNAL DBSJ1,FX,GX,FP1,FQ,DBSJ0,FPP,FQQ,FAA,FBB,FF,GG,FFX,GGX
      EXTERNAL EGX,EG,ERUOX,ERUO,PF,QF,AF,BF
      CALL UMACH (2,NOUT)
      Z1=Z0
```

```
      IF (CONDT) 468,468,469                                            EDD01210
468   Z1=-Z0                                                            EDD01220
469   TEST2=ABS(K0*Z*SQRT(CIT/2.))                                      EDD01230
      IF (TEST2-70.) 001,001,599                                        EDD01240
599   HR1=0.                                                            EDD01250
      HX1=0.                                                            EDD01260
      HR2=0.                                                            EDD01270
      HX2=0.                                                            EDD01280
      GOTO 600                                                          EDD01290
C ----------- CALCULATE H(RUO) OR E(RUO) : HR1,HX1 ---------------------EDD01300
------                                                                  EDD01310
001   TEST1=1.                                                          EDD01320
401   SUM1=0.0                                                          EDD01330
      SUM2=0.0                                                          EDD01340
      TERV=2.                                                           EDD01350
      IRULE=1                                                           EDD01360
      A=0.0                                                             EDD01370
      B=TERV                                                            EDD01380
110   ERRABS=0.0                                                        EDD01390
      ERRREL=1.E-2                                                      EDD01400
      IF (CONDT) 441,441,443                                            EDD01410
443   IF (TEST1) 451,451,453                                            EDD01420
451   CALL INTEG (ERUOX,RESULT)                                         EDD01430
      GOTO 404                                                          EDD01440
453   CALL INTEG (ERUO,RESULT)                                          EDD01450
      GOTO 404                                                          EDD01460
441   IF (TEST1) 402,402,403                                            EDD01470
402   CALL INTEG(FX,RESULT)                                             EDD01480
      GOTO 404                                                          EDD01490
403   CALL INTEG(F,RESULT)                                              EDD01500
404   SUM1=SUM1+RESULT*EXP(K0*Z*FN1((A+B)/2./K0/HH))                    EDD01510
      IF (ABS(SUM1-SUM2)-1.E-20) 111,111,112                            EDD01520
112   A=B                                                               EDD01530
      B=B+TERV                                                          EDD01540
      SUM2=SUM1                                                         EDD01550
      GOTO 110                                                          EDD01560
111   IF (TEST1) 405,405,406                                            EDD01570
406   TEST1=-1.                                                         EDD01580
      HR1=SUM1*Z1/K0/HH                                                 EDD01590
      GOTO 401                                                          EDD01600
405   HX1=SUM1*Z1/K0/HH                                                 EDD01610
      TEST1=1.                                                          EDD01620
      A=0.0                                                             EDD01630
      B=1.                                                              EDD01640
      IRULE=1                                                           EDD01650
410   ERRABS=0.0                                                        EDD01660
      ERRREL=1.E-2                                                      EDD01670
      IF (CONDT) 461,461,463                                            EDD01680
463   IF (TEST1) 471,471,473                                            EDD01690
471   CALL INTEG (EGX,RESULT)                                           EDD01700
      GOTO 413                                                          EDD01710
473   CALL INTEG (EG,RESULT)                                            EDD01720
      GOTO 480                                                          EDD01730
461   IF (TEST1) 411,411,412                                            EDD01740
411   CALL INTEG(GX,RESULT)                                             EDD01750
      GOTO 413                                                          EDD01760
412   CALL INTEG(G,RESULT)                                              EDD01770
480   TEST1=-1.                                                         EDD01780
      HR1=HR1+RESULT*Z1                                                 EDD01790
      GOTO 410                                                          EDD01800
```

```
413   HX1=HX1+RESULT*Z1                                                EDD01810
C ------------ CALCULATE H(Z) OR E(FIA) : HR2,HX2 -----------------    EDD01820
      IF (CONDT) 490,490,491                                           EDD01830
491   AA=RUO                                                           EDD01840
      RUO=0.                                                           EDD01850
490   TEST1=1.                                                         EDD01860
501   SUM1=0.0                                                         EDD01870
      SUM2=0.0                                                         EDD01880
      TERV=2.                                                          EDD01890
      IRULE=1                                                          EDD01900
      A=0.                                                             EDD01910
      B=TERV                                                           EDD01920
310   ERRABS=0.                                                        EDD01930
      ERRREL=1.E-2                                                     EDD01940
      IF (TEST1) 502,502,503                                           EDD01950
502   CALL INTEG(FFX,RESULT)                                           EDD01960
      GOTO 504                                                         EDD01970
503   CALL INTEG(FF,RESULT)                                            EDD01980
504   SUM1=SUM1+RESULT*EXP(K0*Z*FN1((A+B)/2./K0/HH))                   EDD01990
      IF (ABS(SUM1-SUM2)-1.E-30) 311,311,312                           EDD02000
312   A=B                                                              EDD02010
      B=B+TERV                                                         EDD02020
      SUM2=SUM1                                                        EDD02030
      GOTO 310                                                         EDD02040
311   IF (TEST1) 505,505,506                                           EDD02050
506   TEST1=-1.                                                        EDD02060
      HR2=SUM1*Z0/K0/HH                                                EDD02070
      GOTO 501                                                         EDD02080
505   HX2=SUM1*Z0/K0/HH                                                EDD02090
      TEST1=1.                                                         EDD02100
      A=1.0                                                            EDD02110
      B=0.                                                             EDD02120
      IRULE=1                                                          EDD02130
510   ERRABS=0.                                                        EDD02140
      ERRREL=1.E-2                                                     EDD02150
      IF (TEST1) 511,511,512                                           EDD02160
511   CALL INTEG(GGX,RESULT)                                           EDD02170
      GOTO 513                                                         EDD02180
512   CALL INTEG(GG,RESULT)                                            EDD02190
      TEST1=-1.                                                        EDD02200
      HR2=HR2+RESULT*Z0                                                EDD02210
      GOTO 510                                                         EDD02220
513   HX2=HX2+RESULT*Z0                                                EDD02230
600   RETURN                                                           EDD02240
      END                                                              EDD02250
C -------------- FUNCTION FOR CALCULATING H(RUO) --------------------  EDD02260
      REAL FUNCTION F(XT)                                              EDD02270
      REAL X,K0,FN1,FN2,XT                                             EDD02280
      REAL ALOG,EXP                                                    EDD02290
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02300
      COMMON BB,HHZ,RUO0,DD                                            EDD02310
      INTRINSIC  ALOG,EXP,SQRT                                         EDD02320
      X=XT/K0/HH                                                       EDD02330
      F1=(FP(X)*COS(K0*FN2(X)*Z)-FQ(X)*SIN(K0*FN2(X)*Z))               EDD02340
      F1=F1*EXP(-K0*HH*X+K0*(FN1(X)-FN1((A+B)/2./K0/HH))*Z)            EDD02350
      F=F1*DBSJ1(K0*RUO*SQRT(1.+X**2))*DBSJ1(K0*AA*SQRT(1.+X**2))      EDD02360
      RETURN                                                           EDD02370
      END                                                              EDD02380
      REAL FUNCTION FX(XT)                                             EDD02390
      REAL X,K0,FN1,FN2,XT                                             EDD02400
```

```
      REAL ALOG,EXP                                                     EDD02410
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02420
      COMMON BB,HHZ,RUO0,DD                                             EDD02430
      X=XT/K0/HH                                                        EDD02440
      ZKN=K0*Z*FN2(X)                                                   EDD02450
      F1=FP(X)*SIN(ZKN)+FQ(X)*COS(ZKN)                                  EDD02460
      F1=F1*EXP(-K0*HH*X+K0*(FN1(X)-FN1((A+B)/2./K0/HH))*Z)             EDD02470
      ZKN=K0*RUO*SQRT(1.+X**2)                                          EDD02480
      FX=F1*DBSJ1(ZKN)*DBSJ1(K0*AA*SQRT(1.+X**2))                       EDD02490
      RETURN                                                            EDD02500
      END                                                               EDD02510
C ------------ FUNCTION FP(X) ------------------------------------     EDD02520
      REAL FUNCTION FP(X)                                               EDD02530
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02540
      COMMON BB,HHZ,RUO0,DD                                             EDD02550
      F1=UR*X+FN1(X)                                                    EDD02560
      FP=2*UR*X*(FN1(X)*F1+FN2(X)**2)/(F1**2+FN2(X)**2)                 EDD02570
      RETURN                                                            EDD02580
      END                                                               EDD02590
C ------------ FUNCTION FQ(X) ------------------------------------     EDD02600
      REAL FUNCTION FQ(X)                                               EDD02610
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02620
      COMMON BB,HHZ,RUO0,DD                                             EDD02630
      FQ=FN2(X)*UR*X/((UR*X+FN1(X))**2+FN2(X)**2)                       EDD02640
      RETURN                                                            EDD02650
      END                                                               EDD02660
C ------------ FUNCTION FOR INTEGRAL (0.0,1.0) ------------------------ EDD02670
      REAL FUNCTION G(X)                                                EDD02680
      REAL X,ALOG,EXP,K0,FA,FB                                          EDD02690
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02700
      COMMON BB,HHZ,RUO0,DD                                             EDD02710
      INTRINSIC ALOG,EXP,SQRT                                           EDD02720
      ZKN=K0*(HH*X+FM1(X)*Z)                                            EDD02730
      F1=FA(X)*COS(ZKN)+FB(X)*SIN(ZKN)                                  EDD02740
C     F1=F1*EXP(K0*(FM2(X)-FM2((A+B)/2.))*Z)                            EDD02750
      F1=F1*EXP(K0*(FM2(X)*Z))                                          EDD02760
      ZKN=K0*RUO*SQRT(1.-X**2)                                          EDD02770
      G=F1*DBSJ1(ZKN)*DBSJ1(K0*AA*SQRT(1.-X**2))                        EDD02780
      RETURN                                                            EDD02790
      END                                                               EDD02800
      REAL FUNCTION GX(X)                                               EDD02810
      REAL ALOG,BX,EXP,K0,FA,FB                                         EDD02820
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02830
      COMMON BB,HHZ,RUO0,DD                                             EDD02840
      ZKN=K0*(HH*X+FM1(X)*Z)                                            EDD02850
      F1=(FA(X)*SIN(ZKN)-FB(X)*COS(ZKN))*EXP(K0*FM2(X)*Z)               EDD02860
      GX=F1*DBSJ1(K0*RUO*SQRT(1.-X**2))*DBSJ1(K0*AA*SQRT(1.-X**2))      EDD02870
      RETURN                                                            EDD02880
      END                                                               EDD02890
      REAL FUNCTION FA(X)                                               EDD02900
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02910
      COMMON BB,HHZ,RUO0,DD                                             EDD02920
      REAL FM1,FM2                                                      EDD02930
      F1=UR*X+FM1(X)                                                    EDD02940
      FA=2*UR*X*(FM1(X)*F1+FM2(X)**2)/(F1**2+FM2(X)**2)                 EDD02950
      RETURN                                                            EDD02960
      END                                                               EDD02970
      REAL FUNCTION FB(X)                                               EDD02980
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD02990
      COMMON BB,HHZ,RUO0,DD                                             EDD03000
```

```
      REAL FM1,FM2                                                      EDD03010
      F1=UR*X+FM1(X)                                                    EDD03020
      FB=FM2(X)*UR*X/(F1**2+FM2(X)**2)                                  EDD03030
      END                                                               EDD03040
      REAL FUNCTION FM1(X)                                              EDD03050
      REAL FP1                                                          EDD03060
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03070
      COMMON BB,HHZ,RUO0,DD                                             EDD03080
      FM1=SQRT(0.5*(SQRT(FP1(X)**2+CIT**2)+FP1(X)))                     EDD03090
      RETURN                                                            EDD03100
      END                                                               EDD03110
      REAL FUNCTION FM2(X)                                              EDD03120
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03130
      COMMON BB,HHZ,RUO0,DD                                             EDD03140
      REAL FP1                                                          EDD03150
      IF (FP1(X)-CIT*1.E2) 0042,0042,0043                               EDD03160
0043  Y1=CIT/FP1(X)                                                     EDD03170
      FM2=.5*SQRT(CIT*(Y1-.5*Y1**3+.75*Y1**5))                          EDD03180
      GOTO 0044                                                         EDD03190
0042  FM2=SQRT(.5*(SQRT(FP1(X)**2+CIT**2)-FP1(X)))                     EDD03200
0044  RETURN                                                            EDD03210
      END                                                               EDD03220
      REAL FUNCTION FN1(X)                                              EDD03230
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03240
      COMMON BB,HHZ,RUO0,DD                                             EDD03250
      REAL FP2                                                          EDD03260
      FN1=SQRT(.5*(SQRT(FP2(X)**2+CIT**2)+FP2(X)))                      EDD03270
      RETURN                                                            EDD03280
      END                                                               EDD03290
      REAL FUNCTION FN2(X)                                              EDD03300
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03310
      COMMON BB,HHZ,RUO0,DD                                             EDD03320
      INTRINSIC  SQRT                                                   EDD03330
      REAL FP2,K0, SQRT,CIT,APCR,UR                                     EDD03340
C     WRITE(6,*)FP2(X),'----------------FP2'                           EDD03350
      IF (FP2(X)-CIT*1.E2) 0011,0011,0022                               EDD03360
0022  Y1=CIT/FP2(X)                                                     EDD03370
      FN2=.5*SQRT(CIT*(Y1-.5*Y1**3+3./4.*Y1**5))                        EDD03380
      GOTO 0033                                                         EDD03390
0011  FN2=SQRT(.5*(SQRT(FP2(X)**2+CIT**2)-FP2(X)))                     EDD03400
0033  RETURN                                                            EDD03410
      END                                                               EDD03420
      REAL FUNCTION FP2(X)                                              EDD03430
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03440
      COMMON BB,HHZ,RUO0,DD                                             EDD03450
      REAL X,K0,UR,APCR                                                 EDD03460
      FP2=X**2+1-UR*APCR                                                EDD03470
      RETURN                                                            EDD03480
      END                                                               EDD03490
      REAL FUNCTION FP1(X)                                              EDD03500
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03510
      COMMON BB,HHZ,RUO0,DD                                             EDD03520
      FP1=X**2-1+UR*APCR                                                EDD03530
      RETURN                                                            EDD03540
      END                                                               EDD03550
C --------------- FUNCTION FOR CALCULATING H(Z) ---------------------- EDD03560
      REAL FUNCTION FF(XT)                                              EDD03570
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03580
      COMMON BB,HHZ,RUO0,DD                                             EDD03590
      REAL K0                                                           EDD03600
```

```
X=XT/K0/HH                                                          EDD03610
ZKN=K0*FN2(X)*Z                                                     EDD03620
F1=EXP(-K0*HH*X+K0*Z*(FN1(X)-FN1((A+B)/2./K0/HH)))                  EDD03630
F1=(FPP(X)*COS(ZKN)+FQQ(X)*SIN(ZKN))*F1                             EDD03640
FF=F1*DBSJ1(K0*AA*SQRT(1.+X**2))*DBSJ0(K0*RUO*SQRT(1.+X**2))        EDD03650
RETURN                                                              EDD03660
END                                                                 EDD03670
REAL FUNCTION FFX(XT)                                               EDD03680
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03690
COMMON BB,HHZ,RUO0,DD                                               EDD03700
REAL K0                                                             EDD03710
X=XT/K0/HH                                                          EDD03720
ZKN=K0*FN2(X)*Z                                                     EDD03730
F1=FPP(X)*SIN(ZKN)-FQQ(X)*COS(ZKN)                                  EDD03740
F1=F1*EXP(-K0*HH*X+K0*(FN1(X)-FN1((A+B)/2./K0/HH))*Z)               EDD03750
FFX=F1*DBSJ1(K0*AA*SQRT(1.+X**2))*DBSJ0(K0*RUO*SQRT(1.+X**2))       EDD03760
RETURN                                                              EDD03770
END                                                                 EDD03780
REAL FUNCTION GG(X)                                                 EDD03790
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03800
COMMON BB,HHZ,RUO0,DD                                               EDD03810
REAL K0                                                             EDD03820
ZKN=K0*(HH*X+FM1(X)*Z)                                              EDD03830
F1=(FAA(X)*SIN(ZKN)+FBB(X)*COS(ZKN))*EXP(K0*FM2(X)*Z)               EDD03840
GG=F1*DBSJ1(K0*AA*SQRT(1.-X**2))*DBSJ0(K0*RUO*SQRT(1.-X**2))        EDD03850
RETURN                                                              EDD03860
END                                                                 EDD03870
REAL FUNCTION GGX(X)                                                EDD03880
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03890
COMMON BB,HHZ,RUO0,DD                                               EDD03900
REAL K0                                                             EDD03910
ZKN=K0*(HH*X+FM1(X)*Z)                                              EDD03920
F1=(-FAA(X)*COS(ZKN)+FBB(X)*SIN(ZKN))*EXP(K0*FM2(X)*Z)              EDD03930
GGX=F1*DBSJ1(K0*AA*SQRT(1.-X**2))*DBSJ0(K0*RUO*SQRT(1.-X**2))       EDD03940
RETURN                                                              EDD03950
END                                                                 EDD03960
REAL FUNCTION FPP(X)                                                EDD03970
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD03980
COMMON BB,HHZ,RUO0,DD                                               EDD03990
ZKN=UR*X*FN1(X)                                                     EDD04000
FPP=2*UR*X*SQRT(1.+X**2)*ZKN/(ZKN**2+FN2(X)**2)                     EDD04010
RETURN                                                              EDD04020
END                                                                 EDD04030
REAL FUNCTION FQQ(X)                                                EDD04040
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD04050
COMMON BB,HHZ,RUO0,DD                                               EDD04060
FQQ=2*UR*X*SQRT(1.+X**2)*FN2(X)/((UR*X+FN1(X))**2+FN2(X)**2)        EDD04070
RETURN                                                              EDD04080
END                                                                 EDD04090
REAL FUNCTION FAA(X)                                                EDD04100
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD04110
COMMON BB,HHZ,RUO0,DD                                               EDD04120
ZKN=UR*X+FM1(X)                                                     EDD04130
FAA=2*UR*X*SQRT(1.-X**2)*ZKN/(ZKN**2+FM2(X)**2)                     EDD04140
RETURN                                                              EDD04150
END                                                                 EDD04160
REAL FUNCTION FBB(X)                                                EDD04170
COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD04180
COMMON BB,HHZ,RUO0,DD                                               EDD04190
FBB=2.*UR*X*FM2(X)*SQRT(1.-X**2)/((UR*X+FM1(X))**2+FM2(X)**2)       EDD04200
```

```
      RETURN                                                            EDD04210
      END                                                               EDD04220
C ----------   PARAMETERS ! ------------------------                    EDD04230
      SUBROUTINE PARAM(NNF,NNZ,HFIA,HZ,AZZ,AFF,BZZ,BFF)                 EDD04240
C     DIMENSION HFIA(NNF,NNZ),HZ(NNF,NNZ)                               EDD04250
      COMPLEX HFIA(100),HZ(100),AZZ(100,100),AFF(100,100)               EDD04260
      COMPLEX BFF(100,100),BZZ(100,100),Y12,Z12                         EDD04270
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD04280
      COMMON BB,HHZ,RUO0,DD                                             EDD04290
      INTRINSIC SQRT,SIN,COS,ACOS                                       EDD04300
      REAL K1,K0,SQRT,SIN,COS,ACOS                                      EDD04310
      PI=ACOS(-1.)                                                      EDD04320
      U0=4.*PI*1.E-7                                                    EDD04330
      UR=1.                                                             EDD04340
      WRITE(6,*)'INPUT F(HZ)'                                           EDD04350
      READ(5,*)FREQ                                                     EDD04360
      APC0=8.854*1.E-12                                                 EDD04370
      APCR=1.                                                           EDD04380
      WFREQ=2.*PI*FREQ                                                  EDD04390
      K0=WFREQ*SQRT(U0*APC0)                                            EDD04400
      CITA=5.7E7                                                        EDD04410
      CIT=CITA*UR/WFREQ/APC0                                            EDD04420
      AA=.1E-1                                                          EDD04430
      HH=.1E-3                                                          EDD04440
      Z0=K0**2*AA/2.                                                    EDD04450
      Z1=-Z0                                                            EDD04460
      Y12=CMPLX(CITA,WFREQ*APC0*(1+APCR))                               EDD04470
      Z12=CMPLX(0.,-1./WFREQ/U0*(1.+1./UR))                             EDD04480
      BB=1.E-2                                                          EDD04490
      RUO0=2.E-3                                                        EDD04500
      DD=0.                                                             EDD04510
      HHZ=2.E-2                                                         EDD04520
      DO 012 JIZ=0,NNZ                                                  EDD04530
      Z=-(DD+HHZ*JIZ/NNZ)                                               EDD04540
      DO 011 JIF=1,NNF                                                  EDD04550
      JI=JIZ*NNF+JIF                                                    EDD04560
      FIA=2.*PI*JIF/NNF                                                 EDD04570
      RUO=SQRT(RUO0**2+BB**2+2.*RUO0*BB*COS(FIA))                       EDD04580
      CONDT=-1.                                                         EDD04590
      CALL HRUOZ(HR1,HX1,HR2,HX2,CONDT)                                 EDD04600
      CCNS=RUO0*SIN(FIA)/RUO                                            EDD04610
      HFIA(JI)=-(CMPLX(HR1,HX1))*CMPLX(CCNS,0.)                         EDD04620
      HZ(JI)=CMPLX(HR2,HX2)                                             EDD04630
      XXI=BB*COS(FIA)                                                   EDD04640
      YYI=BB*SIN(FIA)                                                   EDD04650
      DO 014 JZ=0,NNZ                                                   EDD04660
      ZZ=-(DD+HHZ*JZ/NNZ)                                               EDD04670
      DO 015 JF=1,NNF                                                   EDD04680
      JII=JZ*NNF+JF                                                     EDD04690
      FIA=2*PI*JF/NNF                                                   EDD04700
      XXJ=BB*COS(FIA)                                                   EDD04710
      YYJ=BB*SIN(FIA)                                                   EDD04720
      RIJ=(XXI-XXJ)**2+(YYI-YYJ)**2+(Z-ZZ)**2                           EDD04730
      IF (RIJ-1.E-10) 017,017,018                                       EDD04740
017   RIJ=SQRT((2*PI*BB/NNF)**2+(HHZ/NNZ)**2)/2.                        EDD04750
      GOTO 019                                                          EDD04760
018   RIJ=SQRT(RIJ)                                                     EDD04770
019   AZZ(JI,JII)=Y12/RIJ-Z12*(1./RIJ**3+3.*(Z-ZZ)**2/RIJ**5)           EDD04780
      AFF(JI,JII)=-3.*(Z-ZZ)/RIJ**5*(Z12)*(XXJ*YYI-XXI*YYJ)             EDD04790
      ALP0=(XXI*XXJ+YYI*YYJ)/RIJ**3+3.*(XXJ*YYI-XXI*YYJ)**2/RIJ**5      EDD04800
```

```
      BFF(JI,JII)=Y12/RIJ-Z12*((ALP0))                                  EDD04810
      BZZ(JI,JII)=-Z12*3.*(XXJ*YYI-XXI*YYJ)*(Z-ZZ)/RIJ**5               EDD04820
C     AZZ(JII,JI)=AZZ(JI,JII)                                           EDD04830
C     AFF(JII,JI)=AFF(JI,JII)                                           EDD04840
C     BFF(JII,JI)=BFF(JI,JII)                                           EDD04850
C     BZZ(JII,JI)=BZZ(JI,JII)                                           EDD04860
015   CONTINUE                                                          EDD04870
014   CONTINUE                                                          EDD04880
011   CONTINUE                                                          EDD04890
012   CONTINUE                                                          EDD04900
      WRITE (6,*)K0,RUO,'-----------------K0 RUO'                       EDD04910
      RETURN                                                            EDD04920
      END                                                               EDD04930
C ---------- FUNCTIONS FOR CALCULATING ERUO -----------------            EDD04940
      REAL FUNCTION ERUO(XT)                                            EDD04950
      REAL X,K0,FN1,FN2,XT                                              EDD04960
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD04970
      COMMON BB,HHZ,RUO0,DD                                             EDD04980
      X=XT/K0/HH                                                        EDD04990
      F1=(PF(X)*COS(K0*FN2(X)*Z)-QF(X)*SIN(K0*FN2(X)*Z))                EDD05000
      F1=F1*EXP(-K0*HH*X+K0*(FN1(X)-FN1((A+B)/2./K0/HH))*Z)             EDD05010
      ERUO=F1*DBSJ0(K0*RUO*SQRT(1.+X**2))                               EDD05020
      RETURN                                                            EDD05030
      END                                                               EDD05040
      REAL FUNCTION ERUOX(XT)                                           EDD05050
      REAL X,K0,FN1,FN2,XT                                              EDD05060
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD05070
      COMMON BB,HHZ,RUO0,DD                                             EDD05080
      X=XT/K0/HH                                                        EDD05090
      ZKN=K0*Z*FN2(X)                                                   EDD05100
      F1=PF(X)*SIN(ZKN)+QF(X)*COS(ZKN)                                  EDD05110
      F1=F1*EXP(-K0*HH*X+K0*(FN1(X)-FN1((A+B)/2./K0/HH))*Z)             EDD05120
      ERUOX=F1*DBSJ0(K0*RUO*SQRT(1.+X**2))                              EDD05130
      RETURN                                                            EDD05140
      END                                                               EDD05150
      REAL FUNCTION EGX(X)                                              EDD05160
      REAL K0                                                           EDD05170
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD05180
      COMMON BB,HHZ,RUO0,DD                                             EDD05190
      ZKN=K0*(HH*X+FM1(X)*Z)                                            EDD05200
      F1=(AF(X)*SIN(ZKN)+BF(X)*COS(ZKN))*EXP(K0*FM2(X)*Z)               EDD05210
      EGX=F1*DBSJ0(K0*RUO*SQRT(1.-X**2))                                EDD05220
      RETURN                                                            EDD05230
      END                                                               EDD05240
      REAL FUNCTION EG(X)                                               EDD05250
      REAL K0                                                           EDD05260
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD05270
      COMMON BB,HHZ,RUO0,DD                                             EDD05280
      ZKN=K0*(HH*X+FM1(X)*Z)                                            EDD05290
      F1=AF(X)*COS(ZKN)-BF(X)*SIN(ZKN)                                  EDD05300
      F1=F1*EXP(K0*FM2(X)*Z)                                            EDD05310
      EG=F1*DBSJ0(RUO*K0*SQRT(1.-X**2))                                 EDD05320
      RETURN                                                            EDD05330
      END                                                               EDD05340
      REAL FUNCTION PF(X)                                               EDD05350
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,QX,QW,NCOEFEDD05360
      COMMON BB,HHZ,RUO0,DD                                             EDD05370
      F1=FN1(X)+APCR*X                                                  EDD05380
      PF=2*X**2*F1/(F1**2+FN2(X)**2)                                    EDD05390
      RETURN                                                            EDD05400
```

```
      END                                                               EDD05410
      REAL FUNCTION QF(X)                                               EDD05420
      QF=-2.*FN2(X)*X**2/((FN1(X)+APCR*X)**2+FN2(X)**2)                 EDD05430
      RETURN                                                            EDD05440
      END                                                               EDD05450
      REAL FUNCTION AF(X)                                               EDD05460
      F1=FM1(X)+APCR*X                                                  EDD05470
      AF=2.*X**2*F1/(F1**2+FM2(X)**2)                                   EDD05480
      RETURN                                                            EDD05490
      END                                                               EDD05500
      REAL FUNCTION BF(X)                                               EDD05510
      BF=2.*X**2*FM2(X)/((FM1(X)+APCR*X)**2+FM2(X)**2)                  EDD05520
      RETURN                                                            EDD05530
      END                                                               EDD05540
C ----------- CALCULATE THE INTEGRAL COEFFICIENTS--------------------  EDD05550
      SUBROUTINE COEF (ICA,QX,QW)                                       EDD05560
      INTEGER II,IWEIGH,NFIX,NOUT                                       EDD05570
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,XX,WW,NCOEFEDD05580
      COMMON BB,HHZ,RUO0,DD                                             EDD05590
      REAL ALPHA,BETA,QXFIX(2),SUMF                                     EDD05600
      REAL QX(21),QW(21)                                                EDD05610
      EXTERNAL GQRUL,UMACH,BAS                                          EDD05620
      CALL UMACH(2,NOUT)                                                EDD05630
      NCOEF=21                                                          EDD05640
      ICA=NCOEF                                                         EDD05650
      IWEIGH=1                                                          EDD05660
      ALPHA=0.                                                          EDD05670
      BETA=0.                                                           EDD05680
      NFIX=0                                                            EDD05690
      CALL GQRUL(NCOEF,IWEIGH,ALPHA,BETA,NFIX,QXFIX,QX,QW)              EDD05700
      RETURN                                                            EDD05710
      END                                                               EDD05720
C ---------- CALCULATE INTEGRAL FROM A TO B ---------------------      EDD05730
      SUBROUTINE INTEG(BAS,SUMF)                                        EDD05740
      EXTERNAL BAS,COEF                                                 EDD05750
      COMMON Z0,WFREQ,U0,APC0,UR,APCR,CIT,AA,HH,K0,RUO,Z,A,B,XX,WW,NCOEFEDD05760
      COMMON BB,HHZ,RUO0,DD                                             EDD05770
      REAL QX(21),QW(21)                                                EDD05780
      CALL COEF(ICA,QX,QW)                                              EDD05790
      SUMF=0.                                                           EDD05800
      DO 020 II=1,ICA                                                   EDD05810
      XINT=(B-A)/2.*QX(II)+(B+A)/2.                                     EDD05820
      SUMF=SUMF+BAS(XINT)*QW(II)                                        EDD05830
020   CONTINUE                                                          EDD05840
      RETURN                                                            EDD05850
      END                                                               EDD05860
```

Research Initiation Program (RIP)

Final Report Submitted to the

Air Force Office of Scientific Research/Universal Energy Systems

Title: Silicon/Tin Polymers for Enhanced Third Order Nonlinear Optical Properties

Principal Investigator: Joseph B. Lambert

Institution: Northwestern University
Evanston, IL 60208

PI Department: Chemistry

Amount: $20,000

Duration: 12 months (November 1, 1990-October 31, 1991)

As described in the original proposal, polysilanes have very favorable non-linear optical properties. In an effort to enhance these properties and possibly obtain values of $\chi^{(3)}$ in a range for practical applications, we have prepared polysilanes that have been doped with the more highly polarizable tin atom. During the grant period we have been successful in obtaining polysilylstannanes, which we now plan to examine for enhanced NLO properties.

The synthesis of polysilanes usually involves condensation of dichlorosilanes with sodium metal. Success of the polymerization reaction often depends on the nature of the solvent, reaction conditions, and the purity of the reactants. Similar procedures have been developed to prepare polystannanes, but chain lengths are much shorter than with polysilanes, presumably because of the much lower bond strength of Sn-Sn. A copolymer of silicon and tin would contain Si-Sn rather than

Sn-Sn bonds and would possibly attain reasonably long chain lengths. Copolymers of Si with Ge are known but not of Si with Sn.

To make poly-Si/Sn we selected the Wurtz-type coupling that has been very successful with polysilanes. We used phenylmethyldichlorosilane ($PhMeSiCl_2$) and dibutyldichlorostannane (Bu_2SnCl_2) as our starting materials. Numerous variants were explored without successful isolation of polymers. Through control reactions, we learned that the Si-Sn bond is unstable to most reaction conditions using metals (typical Wurtz conditions). We finally determined that polymerization of the mixed dichlorides in the presence of lithium metal, with 12-crown-4 as a chelating agent, and in diethyl ether as solvent successfully resulted in polymerization.

The resulting polymer has been characterized by proton, carbon-13, silicon-29, and tin-119 NMR spectroscopy, and by UV spectroscopy. Optimization of reaction conditions with respect to solvent, concentrations, time, and temperature has been carried out, and we have determined that the Si-Sn bond is stable under the reaction conditions. We also have explored changing the ratio of the Si and Sn dichlorides in order to control the Si/Sn ratio in the product. Molecular weights have been examined by mass spectrometry and gel permeation chromatography.

Work is continuing on the preparation, purification, and characterization of these novel materials. In particular, we plan to measure $\chi^{(3)}$ by the third harmonic generation method, which is available in the laboratory of a physics colleague.

CONFORMATIONAL STRUCTURE AND DYNAMICS OF PERFLUOROPOLYALKYLETHER LUBRICANTS

Martin Schwartz
Professor
Department of Chemistry

University of North Texas
215 W. Sycamore
Denton, TX 76203

Final Report For:
Research Initiation Program
Wright Laboratory

Sponsored by:
Air Force Office of Scientific Reserach
Bolling Air Force Base, Washington, D. C.

and

University of North Texas

December 1991

CONFORMATIONAL STRUCTURE AND DYNAMICS OF PERFLUOROPOLYALKYLETHER LUBRICANTS

Martin Schwartz
Professor
Department of Chemistry
University of North Texas

Abstract

In order to establish a methodology for future quantum mechanical investigations of PFPAE model compounds, the geometries, energies and vibrational frequencies of 1,1,2-trichloro-1,2,2-trifluoroethane [TCTFE] were studied by *ab initio* calculations using various basis sets. Bond angles calculated with the 6-31G(d) and 6-311G(d) bases were in close agreement with each other and with experimental results. Comparable energies of the equilibrium conformers and torsional barriers were obtained by MP2 calculations using the 6-31G(d), 6-311G(d) and 6-311G(2df) bases. The calculated equilibrium energy difference is in qualitative agreement with experimental results. It was concluded that the 6-31G(d) basis set provides satisfactory results, comparable to those obtained using larger polarized bases.

Fluorine-19 NMR spin-lattice (T_1) relaxation times were measured as a function of temperature and frequency for several PFPAE's. Derived correlation times and activation energies revealed that there exists a substantially greater barrier to rotation about C-C than C-O bonds in these systems. There is a significant variation in values of E_a obtained at two experimental frequencies (84.7 and 282.2 MHz), indicating that overall and internal polymer chain rotation affect the two measurements differently.

CONFORMATIONAL STRUCTURE AND DYNAMICS OF PERFLUOROPOLYALKYLETHER LUBRICANTS

Martin Schwartz

I. INTRODUCTION

Perfluoropolyalkylether (PFPAE) fluids possess the viscoelastic, thermal and lubricity properties necessary to serve as effective, stable liquid phase lubricants.[1,2] No currently available commercial PFPAE lubricants, however, are capable of operation at the temperature extremes and oxidative conditions required for lubrication of high performance gas turbine engines.

The viscoelastic properties of polymer fluids such as the PFPAE's are, of course, intimately connected to the chain flexibility in these systems which is, in turn, dependent upon the potential energy barriers to internal rotation about single bonds in the polymer. This year, I have been engaged in two research projects designed to obtain a better understanding of chain mobility in PFPAE's. (1) In collaboration with Dr. Harvey L. Paige at the Materials Directorate, Wright Laboratory, Wright-Patterson AFB, I have begun *ab initio* and semi-empirical quantum mechanical modelling studies of rotational barriers in molecules related to perfluoroethers; (2) Here at the University of North Texas, we have been obtaining data on the frequency and temperature dependence of fluorine-19 NMR spin-lattice relaxation times (T_1) in model PFPAE compounds. This latter technique affords a direct measurement of the flexibility at various points along the polymer chain.

In the future, Dr. Paige and I plan to use the results of the quantum mechanical modelling studies in molecular dynamics simulations of the internal rotation in PFPAE chains. The results of these simulations will be compared to the experimental NMR measurments, which

will provide an assessment of the accuracy of the calculations. Ultimately, we expect to be able to use molecular modelling to predict the viscoelastic and conformational properties of these fluids, thus aiding in the molecular design of new PFPAE lubricants.

II. AB INITIO MOLECULAR ORBITAL STUDIES OF ROTATIONAL BARRIERS

Introduction The barriers to internal rotation about the C-O and C-C bonds and, hence, chain flexibility in PFPAE's depend upon both the electronic and steric properties of groups attached to the bonded atoms. Prior to beginning our investigations of the perfluoroethers, Dr. Paige and I decided to establish the methodology *via* studies of two halofluoroethanes, 1,1,2-trichloro-1,2,2-trifluoroethane (CCl_2F-$CClF_2$) [TCTFE] and 1,2-dichloro-1,1,2-trifluoro-2-iodoethane ($CClF_2$-CClFI) [DCTFIE]. The chlorine and iodine atoms in these molecules possess very different steric and electronic properties and provide us with the opportunity to assess the accuracy of the calculational procedures without the added complication of multiple internal rotations. The work on TCTFE has been completed, and an article has been written and accepted for publication in the *Journal of Physical Chemistry*.[3] The computations have been completed on DCTFIE and the analysis is almost complete. In addition, we have measured the Raman vibrational spectra and Dr. James Liang of the Materials Directorate has obtained experimental IR frequency and intensity data on the vibrations in DCTFIE, which will be compared with theoretical predictions. Results on the latter molecule will be presented in an article to be submitted this Spring.[4]

Calculations *Ab initio* molecular orbital calculations were performed using the Gaussian-90[5] program on a Cray X-MP/216 computer.

In TCTFE, the two equilibrium geometries, C_1 and C_s, are termed gauche (G) and trans (T), respectively, to denote the relative positions of the lone fluorine (F_1) on the first carbon atom and the single chlorine (Cl_1) on the second carbon; the transition state structures are called GT and GG′. The equilibrium and saddle point geometries were optimized with the following basis sets: 3-21G,[6] 6-31G,[7] D95,[8] 6-31G(d)[7,9] and 6-311G(d).[9,10] Calculations on the four conformers were also performed with the 6-311G(2df)[9,10] basis set using the 6-311G(d) geometries. Single point second order Møller-Plesset[11] correlation energy calculations were performed with several of the largest basis sets.

For comparison with the *ab initio* results, MNDO,[12] AM1[13] and PM3[14] semi-empirical energy calculations were performed with the program MOPAC.[15] As above, all structural parameters were optimized for both equilibrium and transition state conformations.

Results and Discussion

Geometries Tabulated in columns 3-7 of Table 1 are the structural parameters for the trans conformer calculated with the various basis sets used in this study. The results from the D95 (double-zeta) basis set are not included since, not surprisingly, they are quite similar to those obtained with the 6-31G (doubly-split valence) basis.

One observes that structural parameters determined with the two polarized basis sets, 6-31G(d) and 6-311G(d) are generally quite close; calculated bond lengths are the same to within 0.002-0.006 Å and bond angles differ by an average of <0.4°. It is seen further from the table that $R(CF_2)$,$R(CF_3)$ < $R(CF_1)$, in agreement with the experimental observation[16] that C-F bond lengths shorten with increasing fluorine substitution on a carbon.

Iwasaki[17] has determined the structure of TCTFE by electron diffraction,[18] and his results are given in the second column of Table 1. One finds that C-C and C-Cl bond lengths obtained with the 6-31G(d) and 6-311G(d) basis sets are in close coincidence to his values, whereas calculated C-F bond lengths are significantly shorter than experiment; the latter inequality is common to all large basis set SCF calculations on fluorocarbons.[19] Considering the approximations employed in the experimental structure determination, the bond angles, too, are in quite satisfactory agreement with measured values.

Shown in the last four columns of the table are the structures of both equilibrium and saddle point configurations, determined with the 6-311G(d) basis set. As found in earlier *ab initio* studies of substituted ethanes,[19] all bond lengths with the exception of C-C remain approximately constant, whereas the latter increases by 0.045-0.050 Å in the saddle point structures. One observes, also, that apparently equivalent angles appear to vary within a given conformation and between conformers. These variations can, in all cases, be explained by steric interactions. For instance, the observation that $\angle CCF_2 > \angle CCF_3$ in the G rotamer is due to the fact that fluorine F_2 has two gauche interactions with chlorine atoms compared to only one for F_3. Similarly, $\angle CCF_1$ and $\angle CCF_2$ are greatest in the GG′ transition state, which is the only conformation in which they eclipse chlorine atoms.

Energies Displayed in Table 2 are the energies and energy differences (relative to the T conformer) calculated at the SCF and MP2 levels using the various basis sets. One finds that all *ab initio* calculations yield negative values for the energy difference, ΔE(G-T), which, at the SCF level, ranges from -0.7 to -0.8 kcal/mol for the two

largest basis sets. The correlation energy correction is small, lowering ΔE by 0.0 to 0.2 kcal/mol. There is no further correction to the equilibrium energy difference arising from zero point vibrational energy (ZPVE) and thermal contributions to the enthalpy (discussed in ref. 3). Thus the final calculated range is $\Delta E(G-T) = -0.6$ to -0.7 kcal/mol, which is in qualitative agreement with experimental estimates; $\Delta H(exp) \approx \Delta E(exp) = -0.25$[20] to -0.35[21] kcal/mol.

One sees also from the table that both transition state energies, $\Delta E(GT-T)$ and $\Delta E(GG'-T)$, generally increase with size of the basis set. The correlation energy correction to $\Delta E(GT-T)$ is -0.0 to -0.3 kcal/mol. Together with a -0.4 kcal/mol correction from Δ[ZPVE] and $\Delta[H(T)-H(0)]$ (Ref. 3), one obtains a net energy barrier, $\Delta E(GT-T) = 8.7 - 9.1$ kcal/mol (with the three largest basis sets). This value lies within the range determined from ultrasonic relaxation measurements,[22] $\Delta E(GT-T) = 5.0 - 10.0$ kcal/mol. It is somewhat higher than the estimate from the infrared torsional frequency;[22] however, this latter measurement was subject to potentially large errors.

It is found that the second transition state, GG′, has a significantly lower energy than GT. With the correlation energy, Δ[ZPVE] and [H(T)-H(0)] corrections, $\Delta E(GG'-T) = 6.8 - 7.2$ kcal/mol. The higher energy of the GT saddle point seems quite reasonable since there is a repulsive interaction between two eclipsed chlorine atoms which is not present in the GG′ conformation.

Finally, we note that all three semi-empirical methods[12-14] predict that $\Delta E(G-T) > 0$, which disagrees with both experimental and *ab initio* results. Too, they predict that $\Delta(GT-T) < \Delta(GG'-T)$, in contrast to the *ab initio* calculations and to chemical intuition.

Vibrations Although not shown (see Ref. 3), vibrational frequencies were calculated for the G and T conformers using the 6-31G(d) basis set, and were multiplied by the standard scale factor, 0.90, to account for corrections due to vibrational anharmonicity and electron correlation. The scaled frequencies were in quite satisfactory agreement with experimental data on the two rotamers, with errors of 1.2-1.8% for modes below 1000 cm^{-1}; the error was somewhat greater (3.5%-4.2%) for the C-F stretching vibrations (above 1000 cm^{-1})

Conclusions The agreement between theoretical and experimental geometries, energies and vibrational frequencies were quite satisfactory when using the 6-31G(d) basis set, which provides evidence that this basis should provide an adequate characterization of the structures and energies in our planned computational investigations of model compounds for the perfluoropolyalkylethers.

Planned Investigations Dr. Paige and I have just begun a comparative *ab initio* study of perfluorobutane [$CF_3CF_2CF_2CF_3$] and perfluoroethylmethyl ether [$CF_3CF_2OCF_3$] in order to determine the effect replacement of CF_2 groups by oxygen atoms on the conformational energies and rotational barriers. We plan also to investigate whether the 3-21G(d) basis set yields satisfactory geometric structural parameters. If this proves true, it will permit us to perform more computationally efficient calculations on larger PFPAE model compounds. We also plan to initiate an investigation on the torsional potential energy surface in perfluorodimethoxymethane [$CF_3OCF_2OCF_3$] either later this coming Spring. This molecule is of interest since one expects the torsional barriers to rotation about the the two CF_2-O bonds to be mutually dependent upon both dihedral angles.

III. NMR RELAXATION TIME STUDIES OF PFPAE CHAIN FLEXIBILITY

Introduction The measurement of NMR Spin-Lattice (T_1) relaxation times is a well established technique to probe both the rates and mechanisms of molecular reorientation in liquids and solution.[23] The method has also been used quite profitably to characterize the conformational mobility of flexible chain polymers.[24,25] To date, however, NMR relaxation has not yet been applied to study the polymer chain dynamics in perfluoropolyalkylethers.

As applied to the PFPAE's, the fluorine-19 relaxation time [$T_1(^{19}F)$] is a function of the rotational correlation time (τ_c) of the vector connecting the two fluorine atoms in a CF_2 group; i.e. $T_1^{-1} \propto \tau_c/R^6$, where R is the distance between the two fluorine nuclei. Qualitatively, τ_c is the time it takes for the ^{19}F-^{19}F vector to rotate by one radian ($\approx 60°$). The value of the correlation time, then, provides a direct measure of the polymer's flexibility in the region immediately surrounding a given perfluoromethylene group. For polymeric species such as the PFPAE's, τ_c may be a function of the frequency of the NMR experiment. The acquisition of data at two or more frequencies permits a determination of the mechanism of the internal rotation dynamics in the polymer chain.

During the past year, we have extended investigations begun in Summer, 1990 to study the temperature and frequency dependence of NMR relaxation in a number of perfluoropolyalkylethers. The data acquisition is still in progress; below, we describe the preliminary results obtained to date.

Experimental Fluorine-19 NMR measurements were performed by students in my research group at the University of North Texas on (a) a JEOL

FX90Q FT-NMR Spectrometer operating at B_o=21.1 kG [$\nu_o(^{19}F)$=84.7 MHz], and (b) a Varian VXR-300 FT-NMR operating at B_o=70.5 kG [$\nu_o(^{19}F)$=282.2 MHz]. Spin-lattice relaxation times were determined with the standard Inversion Recovery Fourier Transform (IRFT) pulse sequence,[26] (180°-τ-90°-Acq.), with 10-12 τ values plus $\tau \to \infty$. T_1 was calculated from the peak intensities by a non-linear fit to the three parameter magnetization equation.[27]

The following experiments were performed during the past year: (a) Temperature dependence of relaxation times in perfluoropoly(tetraethylene glycol), $R_fO[(CF_2CF_2O)_4CF_2O]_nCF_3$, R_f=CF_3, CF_3CF_2 (ML088-131) [Table 3];[28] (b) Temperature dependence of $T_1(^{19}F)$ in Fomblin-Z, $R_fO[CF_2O]_m[C_2F_4O]_n[C_3F_6O]_q$, R_f=CF_3, C_2F_5 (ML078-80) [Table 4]; (c) Temperature and frequency dependence of relaxation times in perfluoropoly(ethylene oxide), $CF_3O[CF_2CF_2O]_nCF_3$ (ML088-50) [Table 5]. Correlation times reported in Tables 3-5 were calculated using standard formulae.[29]

Results and Discussion The temperature dependent rotational correlation times of perfluoropoly(tetraethylene glycol) are displayed in Table 3. One notes first that τ_c's of the perfluoromethyl (CF_3) end groups [peaks 4 and 5] are lower than rotational times of perfluoromethylene (CF_2) groups in the middle of the polymer chain. This is to be expected since the internal rotation of the latter groups require cooperative reorientation about several adjacent bonds. One sees, however that the two CF_3 groups are not equivalent; the activation energy for rotation in the CF_3CF_2O- unit [peak 4] is substantially higher than in the CF_3O- group [peak 5] (4.5 versus 2.5 kcal/mol), which provides evidence that, as is intuitively reasonable, the barrier to

rotation about a CF_3-CF_2 bond is substantially greater than for a CF_3-O bond. Further evidence for a lower barrier to rotation about C-O bonds is provided by a comparison of peaks 2, 3 and 6. The first two resonances result from ^{19}F nuclei in CF_2 groups with one C-C and one C-O bond, whereas the third peak represents a group with two C-O bonds. The activation energy for rotation of the latter CF_2 unit is substantially lower (2.9 kcal/mol) than found in the two former peaks (3.8 and 4.1 kcal/mol) which, again, provides evidence for a markedly lower barrier to rotation about the C-O bonds in perfluoropolyalkylethers.

Table 4 contains correlation times obtained for the various resonances of Fomblin-Z as a function of temperature in the neat fluid. One observes the same trend found above in the simpler perfluoroether. Specifically, the activation energies for rotation of CF_2 groups containing both a C-C and C-O bond [peaks 1-4] are uniformly greater than those for perfluoromethylene units bonded to two oxygens [bands 7-9], which likely results from the higher barrier to internal rotation about the former bonds. Due to its greater molecular weight, with consequently greater correlation times, Fomblin-Z is almost assuredly not in the motional narrowing limit. Hence, a complete analysis of the reorientation in this molecule will require measurements of the frequency dependence of its relaxation times, which are currently in progress.

Displayed in Table 5 are the reorientational correlation times of perfluoropoly(ethylene oxide) as a function of temperature at two different fluorine-19 resonance frequencies (84.7 and 282.2 MHz). One observes that, as in perfluoropoly(tetraethylene glycol) and Fomblin-Z, both the values of τ_c and E_a for reorientation of the CF_3O- end group

are lower than for other groups in the middle of the polymer chain. Most significantly, the activation energies of all resonances are markedly less in the high frequency (282.2 MHz) experiments. If all motions in the polymer were at a frequency greater than ~300 MHz, one would expect results of the two experiments to be identical. Hence, it is clear that the low frequency measurements reflect two separate reorientational mechanisms, overall rotation superposed on the slower internal rotations about C-C and C-O bonds in the chain. This coming Spring, we plan to fit the data with several theoretical models of chain reorientation,[30,31] which will permit us to obtain a quantitative separation of the internal and overall polymer dynamics.

<u>Future Studies</u> Work is currently underway to determine the relaxation and correlation times of Fomblin-Z at a second ^{19}F resonance frequency and to study relaxation in two other Fomblins of differing average molecular weights. As for perfluoropoly(ethylene oxide), the data will be fit to models which will permit separation of the rates and mechanisms of internal and overall molecular rotation and will, hopefully, provide semi-quantitative estimates of the barriers to internal rotation; these results will be correlated with the experimental viscosity indices. It is expected that, eventually, these experiments will yield an indication of the effects of structure (e.g. number of CF_2, C_2F_4, C_3F_6 units and their arrangement) on the viscosity-temperature characteristics of perfluoroethers and aid in the design of new PFPAE lubricants.

REFERENCES

1. Snyder, C. E., Jr.; Dolle, R. E., Jr. *ASLE Trans.* **1975**, *19*, 171.

2. Snyder, C. E., Jr.; Gschwender, L. J.; Tamborski, C. *Lubr. Eng.* **1981**, *37*, 344.

3. "*Ab Initio* Study of Conformational Energies and Rotational Barriers in a Chlorofluoroethane," H. L. Paige and M. Schwartz, *J. Phys. Chem.* (in Press).

4. "Vibrational Studies and Ab Initio Calculations on the Conformers of 1,2-Dichloro-1,1,2-trifluoro-2-iodoethane, J. Liang, M. Schwartz and H. L. Paige, *J. Phys. Chem.* (to be Submitted).

5. Gaussian 90, Revision F; Frisch; M. J.; Head-Gordon, M.; Trucks, G. W.; Foresman, J. B.; Schlegel, H. B.; Raghavachari, K.; Robb, M.; Binkley, J. S.; Gonzalez, C.; Defrees, D. J.; Fox, D. J.; Whiteside, R. A.; Seeger, R.; Melius, C. F.; Baker, J.; Martin, R. L.; Kahn, L. R.; Stewart, J. J. P.; Topiol, S.; Pople, J. A.; Gaussian, Inc.: Pittsburgh, PA, 1990.

6. Pietro, W. J.; Francl, M. M.; Hehre, W. J.; Defrees, D. J.; Pople, J. A.; Binkley, J. S. *J. Am. Chem. Soc.* **1982**, *104*, 5039, and references contained therein.

7. (a) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1982**, *56*, 2257; (b) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

8. Dunning, T. H., Jr.; Hay, P. J. In *Methods of Electronic Structure Theory*; Schaefer, H. F., III, Ed.; Plenum Press: New York, 1977; p 1.

9. Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265.

10. (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650; (b) McLean, A. D.; Chandler, G. S. *ibid.* **1980**, *72*, 5639.

11. Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.

12. Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.

13. Dewar, M. J. S.; Zoebisch, E. J.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

14. Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209, 221.

15. Stewart, J. J. P. *MOPAC: Version 5.0*; Frank J. Seiler Research Laboratory, U. S. Air Force Academy, Colorado Springs, Colorado, 80840.

16. (a) Harmony, M. D.; Laurie, V. W.; Kuczkowski, R. L.; Schwendeman, R. H.; Ramsey, D. A.; Lovas, F. J.; Lafferty, W. J.; Maki, A. G. *J. Phys. Chem. Ref. Data* **1979**, *8*, 619; (b) Hirota, E.; Tanaka, T.; Sakakibara, A.; Ohashi, Y.; Morino, Y. *J. Mol. Spectrosc.* **1970**, *34*, 222; (c) Typke, V.; Dakkouri, M.; Oberhammer, H. *J. Mol. Struct.* **1978**, *44*, 85.

17. Iwasaki, M. *Bull. Chem. Soc. Jpn.* **1959**, *91*, (a) 194; (b) 207.

18. In the structure determination of CCl_2F-$CClF_2$, the C-C bond length and the C-C-F and C-C-Cl bond angles were transferred directly from earlier investigations of $CClF_2$-$CClF_2$ and CCl_2F-CCl_2F (Iwasaki, M. *Bull. Chem. Soc. Jpn.* **1958**, *31*, 1071).

19. (a) Dixon, D. A.; Fukunaga, T.; Smart, B. E. *J. Am. Chem. Soc.* **1986**, *108*, 1585, 4027; (b) Dixon, D. A. *J. Phys. Chem.* **1986**, *90*, 2038. (c) Dixon, D. A.; Arduengo, III, A. J. *ibid.* **1987**, *91*, 3195; (d) Dixon, D. A. *ibid.* **1988**, *92*, 86.

20. Braathen, G. O.; Gatial, A.; Klaeboe, P.; *J. Mol. Struct.* **1987**, *157*, 73.

21. Klaeboe, P.; Nielsen, J. R. *J. Mol. Spectrosc.* **1961**, *6*, 379.

22. Pethrick, R. A.; Wyn-Jones, E. *J. Chem. Soc. (A)* **1971**, 54.

23. Boeré, R. T.; Kidd, R. G. *Ann. Rep. NMR Spectrosc.* **1982**, *13*, 319.

24. Heatley, F. *Ann. Rep. NMR Spectrosc.* **1986**, *17*, 189.

25. Hermann, G.; Weill, G. *Macromolecules* **1975**, *8*, 171.

26. Martin, M. L.; Martin, G. J.; Delpuech, J.-J. *Practical NMR Spectroscopy*: Heyden; London, **1980**, Chap. 6.

27. Rodriguez, A. A.; Chen, S. J. H.; Schwartz, M. *J. Magn. Reson.* **1987**, *74*, 114.

28. In addition to the data in Table 3, measurement of the relaxation times in perfluoropoly(tetraethylene glycol) at several lower temperatures are currently in progress.

29. Ref. 26; Chap. 4.

30. Hall, C. K.; Helfand, E. *J. Chem. Phys.* **1982**, *77*, 3275.

31. Viovy, J. L.; Monnerie, L.; Brochon, J. C. *Macromolecules* **1983**, *16*, 1845.

Table 1. Calculated Structural Parameters[a]

Parameter	Experiment[b]	T [3-21G]	T [6-31G]	T [6-31G(d)]	T [6-311G(d)]	G [6-311G(d)]	GT [6-311G(d)]	GG′ [6-311G(d)]
R(CC)	1.54 (0.05)	1.523	1.529	1.549	1.552	1.550	1.599	1.595
$R(CF_1)$	1.38 (0.02_1)	1.363	1.373	1.330	1.324	1.320	1.321	1.319
$R(CF_2)$	1.33 (0.01_4)	1.345	1.360	1.318	1.313	1.311	1.311	1.311
$R(CF_3)$	1.33 (0.01_4)	1.345	1.360	1.318	1.313	1.313	1.311	1.311
$R(CCl_1)$	1.75 (0.02_7)	1.829	1.801	1.749	1.751	1.756	1.756	1.756
$R(CCl_2)$	1.76 (0.02)	1.821	1.808	1.754	1.756	1.761	1.759	1.761
$R(CCl_3)$	1.76 (0.02)	1.821	1.808	1.754	1.756	1.756	1.762	1.761
$\angle CCF_1$	107.1	106.1	105.5	104.9	105.6	107.6	106.6	109.4
$\angle CCF_2$	108 (1.5)	108.7	108.3	107.6	108.0	109.7	107.2	110.3
$\angle CCF_3$	108 (1.5)	108.7	108.3	107.6	108.0	108.3	110.0	110.3
$\angle CCCl_1$	112 (1.5)	112.6	114.6	114.5	114.0	112.1	115.9	112.1
$\angle CCCl_2$	112 (2.0)	111.3	112.0	111.6	111.3	109.5	114.6	111.2
$\angle CCCl_3$	112 (2.0)	111.3	111.9	111.6	111.3	111.4	111.0	111.2
$\angle F_2CF_3$	108.7	109.2	107.7	107.9	107.8	108.4	107.5	107.0
$\angle Cl_2CCl_3$	110.5 (1)	111.6	111.6	111.6	111.4	110.5	109.4	110.0
$\phi(F_1CCCl_1)$	59.5 (1.5)	180.0	180.0	180.0_6	180.0	58.0	120.4	0.0

a) Bond lengths in Angstroms and angles in degrees.
b) Reference 17.

Table 2. Calculated Conformational Energies

Basis set	T	G	GT	GG′
A. Total Energies (Hartrees)				
HF/3-21G	-1743.842 44	-1743.848 03	-1743.831 07	-1743.839 02
HF/6-31G	-1752.247 87	-1752.250 68	-1752.234 13	-1752.238 64
HF/D95	-1752.310 81	-1752.312 21	-1752.296 76	-1752.300 86
HF/6-31G(d)	-1752.458 57	-1752.460 12	-1752.443 56	-1752.447 07
HF/6-311G(d)	-1752.630 64	-1752.631 75	-1752.615 42	-1752.618 83
HF/6-311G(2df) //6-311G(d)	-1752.673 92	-1752.675 17	-1752.658 76	-1752.661 81
MP2/6-31G(d)	-1753.624 56	-1753.626 06	-1753.610 05	-1753.613 10
MP2/6-311G(d)	-1753.933 15	-1753.934 24	-1753.917 97	-1753.921 11
MP2/6-311G(2df) //6-311G(d)	-1754.277 79	-1754.278 75	-1754.263 01	-1754.265 66
B. Relative Energies (kcal/mol)				
HF/3-21G	0.00	-3.50	+7.14	+2.15
HF/6-31G	0.00	-1.76	+8.62	+5.79
HF/D95	0.00	-0.88	+8.82	+6.25
HF/6-31G(d)	0.00	-0.97	+9.42	+7.22
HF/6-311G(d)	0.00	-0.69	+9.55	+7.41
HF/6-311G(2df) //6-311G(d)	0.00	-0.79	+9.51	+7.60
MP2/6-31G(d)	0.00	-0.94	+9.11	+7.19
MP2/6-311G(d)	0.00	-0.68	+9.52	+7.55
MP2/6-311G(2df) //6-311G(d)	0.00	-0.61	+9.27	+7.61
MNDO	0.00	+0.82	+5.17	+7.23
AM1	0.00	+1.13	+5.15	+9.21
PM3	0.00	+0.81	+2.86	+3.44
Experiment	0.00	-0.25[a] -0.35[d]	+5.9[b] +5-10[c]	

a) Reference 20.
b) Reference 22. From torsional frequency (IR).
c) Reference 22. From ultrasonic relaxation.
d) Reference 21.

Table 3. NMR Correlation Times in Perfluoropoly(tetraethylene glycol), $R_fO[(CF_2CF_2O)_4CF_2O]_nCF_3$, R_f=CF_3, CF_3CF_2 (ML088-131) Bulk Fluid. B_o = 21.1 kG [$\nu_o(^{19}F)$ = 84.7 MHz]

Peak	Delta[a]	Assignment	23°C	49°C	73°C	$-E_a$
1	73.6 ppm	$CF_3OCF_2CF_2O$-	72 ps	41 ps	32 ps	3.3 kcal/mol
2	74.0	-$OCF_2CF_2OCF_2O$-	115	63	46	3.8
3	75.7	-$OCF_2CF_2OCF_2CF_2O$-	125	68	46	4.1
4	76.8	CF_3CF_2O-	31	20	10	4.5
5	108.2	$CF_3OCF_2CF_2O$-	35	23	19	2.5
6	112.6	-$OC_2F_4OCF_2OC_2F_4O$-	92	67	45	2.9

a) Chemical shifts are measured in ppm downfield from hexafluorobenzene

Table 4. NMR Correlation Times in Fomblin-Z,
$R_fO[CF_2O]_m[C_2F_4O]_n[C_3F_6O]_q$, R_f=CF_3, C_2F_5 (ML078-80)
Bulk Fluid. B_o = 21.1 kG [$\nu_o(^{19}F)$ = 84.7 MHz]

Peak	Delta[a]	Assignment	-30°C	-3°C	23°C	50°C	76°C	$-E_a$[b]
1	74.2	$OCF_2\mathbf{CF_2}OCF_2O$	337	180	101	60	42	3.4
2	75.9	$OCF_2\mathbf{CF_2}OC_2F_4O$	419	247	139	79	56	3.3
3	79.5	$OCF_2CF_2\mathbf{CF_2}OCF_2O$	420	171	85	59	35	3.9
4	81.1	$OCF_2CF_2\mathbf{CF_2}OC_2F_4O$	371	139	67	52	16	4.6
5	106.9	$\mathbf{CF_3}OCF_2O$	32	22	18	17	16	1.1
6	108.6	$\mathbf{CF_3}OC_2F_4O$	72	42	21	9	7	3.9
7	109.5	$OCF_2\mathbf{OCF_2}OCF_2O$	197	115	67	45	36	2.8
8	111.3	$OCF_2\mathbf{OCF_2}OC_2F_4O$	284	162	97	60	45	3.0
9	112.9	$OC_2F_4\mathbf{OCF_2}OC_2F_4O$	370	215	132	76	60	3.0

a) Chemical shifts are measured in ppm downfield from hexafluorobenzene

Table 5. NMR Correlation Times in Perfluoropoly(ethylene oxide) , $CF_3O[CF_2CF_2O]_nCF_3$ (ML088-50)

5.A Bulk Fluid. B_o = 21.1 kG [$\nu_o(^{19}F)$ = 84.7 MHz]

Peak	Delta[a]	Assignment	10°C	23°C	40°C	55°C	70°C	$-E_a$
1	73.6 ppm	$CF_3OCF_2CF_2O$-	115 ps	79 ps	51 ps	45 ps	30 ps	4.1 kcal/mol
2	75.7	$CF_3CF_2OCF_2CF_2O$-	218	163	111	87	58	4.1
3	75.8	-OCF_2CF_2O-	277	214	137	103	72	4.3
4	76.9	$CF_3CF_2OCF_2CF_2O$-	100	72	48	34	25	4.5
5	108.1	$CF_3OCF_2CF_2O$-	48	42	29	26	22	2.6

5.B Bulk Fluid. B_o = 70.5 kG [$\nu_o(^{19}F)$ = 282.2 MHz]

Peak	Delta[a]	Assignment	10°C	23°C	40°C	55°C	70°C	$-E_a$
1	73.6 ppm	$CF_3OCF_2CF_2O$-	127 ps	86 ps	71 ps	55 ps	43 ps	3.3 kcal/mol
2	75.7	$CF_3CF_2OCF_2CF_2O$-	184	134	115	90	67	3.1
3	75.8	-OCF_2CF_2O-	225	196	167	134	106	2.4
4	76.9	$CF_3CF_2OCF_2CF_2O$-	92	62	48	35	29	3.7
5	108.1	$CF_3OCF_2CF_2O$-	41	34	28	24	21	2.1

a) Chemical shifts are measured in ppm downfield from hexafluorobenzene
b) Activation energies are given in kcal/mol

1990 USAF/UES Research Initiation Program

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

MODELING OF THE FORMATION OF MACROSEGREGATION DURING CASTING SOLIDIFICATION

Prepared by :	Hai-Lung Tsai
Academic Rank:	Assistant Professor
Department and	Mech. and Aero. Engr. and Engr. Mech.
University:	University of Missouri-Rolla

Date:	December 31, 1991
Contract No:	F49620–88–C–0053/SB5881–0378

MODELING OF THE FORMATION OF MACROSEGREGATION DURING CASTING SOLIDIFICATION

by

Hai-Lung Tsai

ABSTRACT

A mathematical model for prediction the formation of macrosegregation in castings has been successfully developed. The fluid flow caused by temperature gradients, solutal gradients, and shrinkage, as well as the domain change during solidification were considered in the model. A benchmark problem of unidirectional solidification chilling from bottom was solved, for the first time, by the comprehensive mathematical model, including the calculations of heat transfer, fluid flow, and mass transfer. The predicted results for the solidification of Al-Cu alloys were consistent with the experimental results reported in 1960s. Some very interesting transient dynamic behaviors for the formation of macrosegregation were discovered by the model, which cannot be obtained from experiments. It was found from the present study that the formation of macrosegregation in castings can be significantly influenced by the fluid flow caused by shrinkage. It is expected that the results obtained from the sponsored research will lead to two journal publications, which are currently under preparation.

I. INTRODUCTION

A large scale of nonuniformity in compositions, i.e., macrosegregation, has frequently been found in large solidified ingots [1-2]. The capability to obtain an ingot with uniform composition is necessary in order to guarantee the quality of subsequent cast parts using the ingots as the raw material. Even for small cast parts, the requirement of high quality parts used, for example, in the aircraft, cannot allow a nonuniformity in composition, because the mechanical properties of the cast part could be greatly influenced by the existence of macrosegregation. Hence, to understand the fundamental mechanisms responsible for the formation of macrosegregation and to find possible methods to prevent or reduce macrosegregation, are the long-term desire in the foundry industry.

The formation of macrosegregation in castings has been understood to be caused by the fluid flow in the casting [3-4]. Fluid flow in the casting can be caused by density difference between the solid and the liquid phases (shrinkage induced flow), as well as the thermal and solutal gradients (natural convection). The modeling of fluid flow and heat transfer in a casting is inherently very difficult due to the presence of solid phase, mushy zone, and liquid phase, as well as the release of latent heat at the unknown solid-liquid interfaces. Recently, the development of continuum model or volume averaging method [5] has made the modeling of casting solidification with fluid flow possible. In the continuum model, only a set of governing equations is required for the entire domain, including the solid phase, mushy zone, and liquid phase.

Previous studies on the formation of macrosegregation in castings were primary concentrated on experimental work. The first attempt to model the formation of macrosegregation is described in the pioneering papers of Flemings and co-workers [6-8], who computed interdendritic flows by treating the mushy zone as a porous medium for

which porosity varied with solid fraction. Solutal segregation in the mushy region was predicted, subject to a prescribed flow which was assumed to be known. A major extension of the model was made by Mehrabian et al. [9], who included the calculations of fluid flow in the mushy zone. Darcy's law was used to describe two-dimensional velocity and pressure fields, although temperature gradients and solidification rate were prescribed and the average concentration of the solid and liquid phases was assumed to be constant. Although their model only allowed for liquid density variations with temperature, Szekely and Jassal [10] appear to have been the first to solve the two-dimensional momentum and energy equations for both the mushy zone and the bulk liquid. From their research, the important conclusion was that the convection in the bulk liquid had a marked effect on the velocity field in the mushy zone. Bennon and Incropera [11] predicted channel segregation of an NH_4Cl–H_2O system by solving the fully coupled momentum, energy, and species equations. A similar model was also established by Beckermann and Viskanta [12]. In their models, the fluid flow was caused by temperature and solutal gradients, but the shrinkage effect was neglected and a constant density was assumed throughout the entire domain.

For many alloys, the density of the solid phase is greater than the density of the liquid phase. Hence, most alloys shrinkage during solidification by 2-8% in volume. The shrinkage inevitably results in fluid flow and global domain change, which are naturally present in phase-change systems and usually cannot be avoided. In practical foundry operations, risers or hot tops are used to feed the shrinkage and thus to reduce possible casting defects. Although the shrinkage induced fluid flow is expected to be small compared with a typical flow in forced convection, Chiang and Tsai [13-14] have found that, depending upon solidification conditions, the shrinkage induced fluid flow can be comparable to or of the same order of magnitude as natural convection caused by thermal

and/or solutal gradients. This is especially true for the fluid flow in the mushy zone. As the fluid flow in the mushy zone has been recognized as the major cause for the formation of macrosegregation, it is expected that the consideration of shrinkage effect is essential for the prediction of macrosegregation in castings. In the present study, a mathematical model to predict the formation of macrosegregation in castings will be established, which includes the natural convection caused by temperature and solutal gradients, as well as the fluid flow caused by shrinkage.

II. FORMULATION OF THE PROBLEM

The formulation based on the continuum model developed by Chiang and Tsai [13-14] will be used. In the derivation of the governing equations, the following assumptions are made: 1) The solid-liquid in the mushy zone are in local thermal and phase equilibrium; 2) Shrinkage can be caused by phase change from the liquid to the solid phases; 3) The Boussinesq approximation can be employed for the natural convection; and 4) The feeding to shrinkage by the gravitational force is complete so that there are no pores inside the solid and the liquid phases. Two solidification conditions will be studied, the first one is a unidirectional solidification chilling from bottom (Figure 1) and the other is a solidification process starting from the side-wall (Figure 2). In the first case, for the solidification of Al-Cu alloys, there is no flow in the casting if shrinkage effect is neglected. In the second case, the fluid flow is caused by both natural convection and shrinkage, as well as their interaction. In the following, the continuum properties are first defined, then the governing equations are given. The meaning of each variable is provided in the nomenclature.

Definitions:

$$\rho = g_s \rho_s + g_l \rho_l ; \quad c = f_s c_s + f_l c_l ; \quad k = g_s k_s + g_l k_l$$

$$D = f_s D_s + f_l D_l ; \quad f_s = \frac{g_s \rho_s}{\rho} ; \quad f_l = \frac{g_l \rho_l}{\rho}$$

$$\mathbf{V} = f_s \mathbf{V}_s + f_l \mathbf{V}_l ; \quad h = f_s h_s + f_l h_l ; \quad f^\alpha = f_s f_s^\alpha + f_l f_l^\alpha \tag{1}$$

Continuity:

$$\frac{\partial}{\partial t}(\rho) + \nabla \cdot (\rho \mathbf{V}) = 0 \tag{2}$$

Momentum:

$$\frac{\partial}{\partial t}(\rho u) + \nabla \cdot (\rho \mathbf{V} u) = \nabla \cdot (\mu_l \frac{\rho}{\rho_l} \nabla u) - \frac{\partial p}{\partial x} - \frac{\mu_l}{K} \frac{\rho}{\rho_l} (u - u_s)$$

$$- \frac{C \rho^2}{K^{1/2} \rho_l} |u - u_s| (u - u_s) - \nabla \cdot (\rho f_s f_l \mathbf{V}_r u_r) + \nabla \cdot (\mu_l u \nabla(\frac{\rho}{\rho_l})) \tag{3}$$

$$\frac{\partial}{\partial t}(\rho v) + \nabla \cdot (\rho \mathbf{V} v) = \nabla \cdot (\mu_l \frac{\rho}{\rho_l} \nabla v) - \frac{\partial p}{\partial y} - \frac{\mu_l}{K} \frac{\rho}{\rho_l} (v - v_s)$$

$$- \frac{C \rho^2}{K^{1/2} \rho_l} |v - v_s| (v - v_s) - \nabla \cdot (\rho f_s f_l \mathbf{V}_r v_r) + \nabla \cdot (\mu_l v \nabla(\frac{\rho}{\rho_l}))$$

$$+ \rho g (\beta_T (T - T_o) + \beta_s (f_l^\alpha - f_{l,o}^\alpha)) \tag{4}$$

Energy:

$$\frac{\partial}{\partial t}(\rho h) + \nabla \cdot (\rho \mathbf{V} h) = \nabla \cdot (\frac{k}{c_s} \nabla h) + \nabla \cdot (\frac{k}{c_s} \nabla (h_s - h)) - \nabla \cdot (\rho (\mathbf{V} - \mathbf{V}_s)(h_l - h)) \tag{5}$$

Species:

$$\frac{\partial}{\partial t}(\rho f^\alpha) + \nabla \cdot (\rho \mathbf{V} f^\alpha) = \nabla \cdot (\rho D \nabla f^\alpha) + \nabla \cdot (\rho D \nabla (f_l^\alpha - f^\alpha)) - \nabla \cdot (\rho (\mathbf{V} - \mathbf{V}_s)(f_l^\alpha - f^\alpha)) \tag{6}$$

The above governing equations are "closed" by providing the relationships between the enthalpy and the temperature, between the solutal concentration and the temperature through the phase diagram, as well as the permeability function in the mushy zone. These auxiliary equations are given as follows:

$$h_s = c_s T; \quad h_l = c_l T + (c_s - c_l) T_e + H; \quad f_s = \frac{1}{1 - k_p}\left[\frac{T - T_l}{T - T_m}\right]$$

$$f_s^\alpha = \left[\frac{k_p}{1 + f_s (k_p - 1)}\right] f^\alpha; \quad f_l^\alpha = \left[\frac{1}{1 + f_s (k_p - 1)}\right] f^\alpha$$

$$K = \frac{g_l^3}{c_1 (1 - g_l)^2}; \quad c_1 = \frac{180}{d^2}; \quad C = 0.13\, g_l^{3/2} \tag{7}$$

The corresponding boundary conditions should be consistent with the conditions given in either Figure 1 or Figure 2. The initial casting temperature is assumed uniform.

III. COMPUTATIONAL CONSIDERATIONS

The above governing equations are in the general format suggested by Patankar [15] for the numerical solution of heat transfer and fluid flow problems; i.e., they contain a transient term, a diffusive term, and source terms. Hence, any established numerical procedure for solving coupled elliptic partial differential equations can be used, with slight modifications for different source terms. In the present study, the equations were solved iteratively at each time step using the control-volume-based finite difference procedure described by Patankar. A fully implicit formulation was used for the time-dependent terms and the combined convection/diffusion coefficients were evaluated using an upwind scheme. The SIMPLEC algorithm [16] was applied to solve the momentum and continuity equations to obtain the velocity field.

At each time step, the momentum equations were solved first in the iteration process, using the estimated volume fraction of solid and liquid for the mixture density, the permeability, and the mass fraction of solid and liquid. Then the energy equation was solved to obtain enthalpy with which the temperature can be calculated. Similarly, the species equation was solved for the solutal concentration. Next, the volume fraction of solid and liquid, the permeability, and the mass fraction of solid and liquid were updated, and this process was repeated for each iteration. For each time step, iterations were terminated when the maximum residual source of mass, momentum, energy, and concentration was less than 1×10^{-5}. A line-by-line solver based on the tridiagonal matrix algorithm (TDMA) was used to iteratively solve the algebraic discretization equations. The last five terms on the right-hand side of equation (3), the last seven terms on the right-hand side of equation (4), and the last two terms on the right-hand side of equations (5) and (6) represent source terms, and they were treated according to the procedure outlined in

Partankar's book [15]. The last two terms on the right-hand side of equation (3), the fifth and sixth terms on the equations (4), as well as the last term on the right-hand side of equations (5) and (6) were calculated via the upwind scheme.

As the domain is changed during solidification, special attention should be drawn to the moving boundary on the free surface at the top of the riser. Moving nodes were placed at the free surface to track the physical domain of the riser. As the free surface was assumed to be flat, the movement of the free surface was handled by taking the averaged velocity of the moving nodes times the time-step size. Detailed description about the free surface can be found in Reference [17].

From numerical experimentation, it was found that a larger time-step size decreased the convergence rate at the beginning of the computation. This is understandable, as the heat flux at the walls is very large at the beginning of solidification. In order to obtain optimum solution accuracy and to maintain numerical stability, a variable time-step size was adopted in the numerical calculation. The numerical algorithm used in the present study has been verified by applying it to several well-known problems where the analytical solutions and/or numerical solutions are available for comparison [17]. The initial time-step size is 0.0001 seconds and the maximum time-step size is 0.05 seconds. All the calculations were executed in Apollo 10,000 workstations.

IV. RESULTS AND DISCUSSION

The results for two solidification conditions as described before will be separately presented in the following.

Case I: Unidirectional Solidification Chilling from Bottom

Figure 3 shows the flow field in the solidifying casting at time t = 16.2 seconds. For Al-Cu alloys, as the density of copper is greater than that of aluminum and the cooling is from bottom, both the temperature and concentration fields are stable. In other words, there is no fluid flow if the shrinkage effect is neglected. Hence, the fluid flow as shown in Figure 3 is caused solely by shrinkage. The corresponding isotherms are given in Figure 4, which indicate a constant temperature distribution in the horizontal direction. Similarly, a uniform distribution of solute (copper) is also found in the horizontal direction. Another set of velocity profile and temperature distribution are given, respectively, in Figure 5 and Figure 6, at time t = 136.2 seconds. The velocity in the pure liquid phase appears to be a parabolic shape, and the velocity profile is distorted when the flow approaches the liquidus line at which the solid phase starts to present. The flow in the mushy zone tends to become uniform in the horizontal direction and its magnitude decreases as the solid fraction increases toward the bottom surface.

The copper concentration distribution along the y-direction at time t = 16.2 seconds is shown in Figure 7. The maximum concentration occurs at the bottom wall and the concentration above the liquidus line is uniform, which is the same as the initial copper concentration, 4.1%. The concentration profile propagates in the y-direction with time, as shown in Figure 8. The copper concentration near the bottom wall is increased with time. It is noted that, from the concentration distribution in both Figures 7 and 8, the concentrations below and above the initial value are compensated each other, so that the overall conservation of mass (copper) is maintained. The dynamic evolution of copper concentration in the y-direction is shown in Figure 9. It is noted that the concentration in the mushy zone is increased with time until the alloy is completely solidified, and then the concentration in the solidified alloy remains constant. The predicted final concentration

distribution in the solidified alloy, along with published experimental results, is given in Figure 10. For comparison, previously calculated results are also given in the figure [18-19]. Although similar experimental and theoretical results were also reported in 1960s by Flemings and co-workers [6-8], a direct comparison was not possible because the exact casting conditions were not given in References [6-8]. Nevertheless, all the previous experiments showed the formation of "inverse segregation" at the bottom of the casting. It is noted that the previously calculated results were obtained under very simplified conditions, i.e., assuming the fluid flow and the velocity of the isotherm were known [6-8,19]. Figure 10 indicates that the model established by the present study has correctly predicted the formation of macrosegregation. Furthermore, the present study proves that the macrosegregation for the case of unidirectional solidification of Al-Cu alloys cooling from bottom is created solely by the shrinkage induced fluid flow.

Case II: Solidification Starting from the Side-Wall

Figure 11 shows the flow pattern in the casting at time t = 9.1 seconds. A clockwise circulation vortex is formed due to the natural convection caused by temperature gradients. The two lines in the casting represent, respectively, the liquidus line and the solidus line. The riser in the figure provides liquid metal to feed shrinkage in the casting. The corresponding isotherms are given in Figure 12, showing the distorted isotherms due to the fluid flow. However, the isotherms closer to the side-wall are not distorted as much as those further from the side-wall. Figure 12 also shows that at time t = 9.1 seconds, the large portion of the casting domain is not affected and still remains at the initial temperature, although a large circulation vortex exists. The corresponding concentration profile is given in Figure 13 at time t = 9.1 seconds. In the solid phase (can be seen in Figure 11), the copper concentration is greater near the wall, and then decreases in the direction away from

the wall. However, the copper concentration in the mushy zone becomes irregular and a concentration "cell" appears.

Another set of flow pattern, isotherms, and concentration profile are given, respectively, in Figures 14, 15, and 16, at time t = 57.1 seconds. The strength of the circulation vortex becomes weaker and the size of the mushy zone increases. The isotherms in Figure 15 are nearly vertical, which implies that the fluid flow does not significantly affect the temperature field. This is because the high thermal conductivity of the Al-Cu alloys and the formation of solid phase which increases the thermal resistance in the casting. Very complex concentration profile, involving several cells, is found in Figure 16. By comparing Figures 13 and 16, it is seen that the concentration profile near the side-wall is similar to that found in the unidirectional solidification at which the flow is caused by shrinkage only. At the beginning of solidification, the fluid flow caused by shrinkage is the strongest due to the large heat flux, while the natural convection is weak. Hence, the concentration near the side-wall found in Figures 13 and 16 are similar to the phenomena observed in Figure 10. However, at later times, the natural convection becomes dominate and overcomes the shrinkage induced fluid flow and, as a result, the concentration profile away from the side-wall is quite different. A large cell of concentration is revealed each in the solid phase and in the mushy zone. As no experimental results are available for comparison, it is not clear that the calculated concentration profile is correct or not. However, the velocity flow pattern and the isotherms in the casting given in Figures 14 and 15 appear correct. As the velocity, temperature, and concentration are coupled in the mathematical model, if the results of temperature and velocity fields are correct, it is a reasonable assumption that the concentration profile predicted by the present model is correct, at least its trend. However, it is feel that corresponding experiments should be conducted to verify the theoretical predictions.

V. CONCLUSIONS

A mathematical model based on the continuum approach was developed for the prediction of macrosegregation in castings during alloy solidification. In the model, the fluid flow caused by temperature gradients, solutal gradients, and shrinkage, as well as the domain change were considered. The set of governing equations are valid for the entire domain, including the solid phase, mushy zone, and liquid phase. Two solidification cases were studied: the first one is a unidirectional solidification chilling from below, while the other is a solidification process starting from side-wall. For the first case, the experimental results for Al-Cu alloys were reported in 1960s and for the first time the present study was able to predict these results. This is because, for the solidification cooling from bottom, the flow is caused only by shrinkage and the existing models have neglected shrinkage effect.

When a chill is placed at the side-wall, both the natural convection and the shrinkage induced fluid flow exist, and they interact each other. The final solute distribution in the solidified casting depends on the relative strength of these two flow mechanisms. As the correct values of several parameters, such as the solutal and thermal expansion coefficients, are not available, numerical experiments were conducted. It was found that under different conditions, the predicted temperature and velocity fields are similar, while the concentration profiles are very much different. This implies that the distribution of solute in the casting is very sensitive to the casting conditions. Corresponding experiments should be conducted to verify the theoretical predictions by the present model. It is expected that the results from the sponsored research will lead to two journal publications, which are currently under preparation.

ACKNOWLEDGEMENTS

The author wish to thank the Air Force System Command and the Air Force Office of Scientific Research for sponsorship of this research. Thanks are also extended to Universal Energy System for their concern and help to me in all administrative and directional aspects of this program.

The concern and support of Jim Malas through the course of the research was greatly appreciated. Stimulating discussions with Bill O'Hara, Carl Lombard, and Venkat Seetharaman were really helpfully.

NOMENCLATURE

c	specific heat
C	coefficient in momentum equation
c_1	permeability coefficient
d	dendrite arm spacing
D	mass diffusion coefficient
f	mass fraction
F	fractional volume of the cell
g	volume fraction or gravitational accel.
h	enthalpy
H	latent heat
k	thermal conductivity
K	permeability
k_p	equilibrium partition ratio
p	pressure
T	temperature
t	time
u, v	x-, y-direction velocity components
V	velocity vector
x, y	Cartesian Coordinates

Greek Symbols

β_s	solutal expansion coefficient
β_T	thermal expansion coefficient
μ	viscosity
ρ	density

Subscripts

e	eutectic
l	liquid
m	fusion temperature, $f^\alpha = 0$
o	reference
r	solid & liquid relative velocity
s	solid

Superscript

α	constituent

REFERENCES

1. M. R. Bridge and J. Beech, "Formation of Macrosegregation in Large, As-cast Masses: Direct Observations of Some Aspects", in Solidification Technology in the Foundry and Cast House, The Metall. Society, pp. 478-483, 1980.

2. P. C. Morgan, P. W. Waterworth, and I. G. Davies, "Macrosegregation in Killed Steel Ingots", in Solidification Technology in the Foundry and Cast House, The Metall. Society, pp. 450-460, 1980.

3. R. J. McDonald and J. D. Hunt, "Fluid Motion Through the Partially Solid Regions of a Casting and Its Importance in Understanding A-type Segregation", Trans. of AIME, Vol. 245, pp. 1993-1997, 1969.

4. M. C. Flemings, Solidification Processing, McGraw Hill, New York, 1974.

5. W. D. Bennon and F. P. Incropera, "A Continuum Model for Momentum, Heat and Species Transport in Binary Solid-Liquid Phase Change Systems I. Model Formulation", Int. J. Heat Mass Transfer, Vol. 30, No. 10, pp. 2161-2170, 1987.

6. M. C. Flemings and G. E. Nereo, "Macrosegregation: Part I", The Metall. Society of AIME, Vol. 239, pp. 1449-1661, 1967.

7. M. C. Flemings, R. Mehrabian, and G. E. Nereo, "Macrosegregation: Part II", The Metall. Society of AIME, Vol. 242, pp. 41-49, 1968.

8. M. C. Flemings and G. E. Nereo, "Macrosegregation: Part III", The Metall. Society of AIME, Vol. 242, pp. 50-55, 1968.

9. R. Mehrabian, M. Keane, and M. C. Flemings, "Interdendritic Fluid Flow and Macrosegregation: Influence of Gravity", Metall Trans., Vol. 1, pp. 1209-1220, 1970.

10. J. Szekely and A. S. Jassal, "An Experimental and Analytical Study of the Solidification of a Binary Dendrite System", Metall. Trans B, 9B, pp. 389-398, 1978.

11. W. D. Bennon and F. P. Incropera, "A Continuum Model for Momentum, Heat and Species Transport in Binary Solid-Liquid Phase Change Systems II. Application to Solidification in a Rectangular Cavity", Int. J. Heat Mass Transfer, Vol. 30, No. 10, pp. 2171-2187, 1987.

12. C. Beckermann and R. Viskanta, "Double-Diffusive Convection During Dendritic Solidification of a Binary Mixture", PCH PhysicoChemical Hydrodynamics, Vol. 10, pp. 195-213, 1988.

13. K. C. Chiang and H. L. Tsai, "Shrinkage-Induced Fluid Flow and Domain Change in Two-Dimensional Alloy Solidification", Int. J. Heat Mass Transfer (to appear).

14. K. C. Chiang and H. L. Tsai, "Interaction between Shrinkage-Induced Fluid Flow and Natural Convection During Alloy Solidification", Int. J. Heat and Mass Transfer (to appear).

15. S. V. Patankar, Numerical Heat Transfer and Fluid Flow, Hemisphere, New York, 1980.

16. J. P. V. Doormaal and G. D. Raithby, "Enhancement of the SIMPLE Method for Predicting Incompressible Fluid Flows", Numerical Heat Transfer, Vol. 7, pp. 147-163, 1984.

17. K. C. Chiang, "Studies on the Shrinkage-Induced Transport Phenomena During Alloy Solidification", Ph.D. Dissertation, University of Missouri-Rolla, 1990.

18. I. Ohnaka, "Microsegregation and Macrosegregation", in Metals Handbook, Vol. 15 Casting, Ninth Edition, pp. 136-141, 1988.

19. H. Kato and J. R. Cahoon, "Inverse Segregation in Directionally Solidified Al-Cu-Ti Alloys with Equiaxed Grains", Metall. Trans A, Vol. 16A, pp. 579-587, 1985.

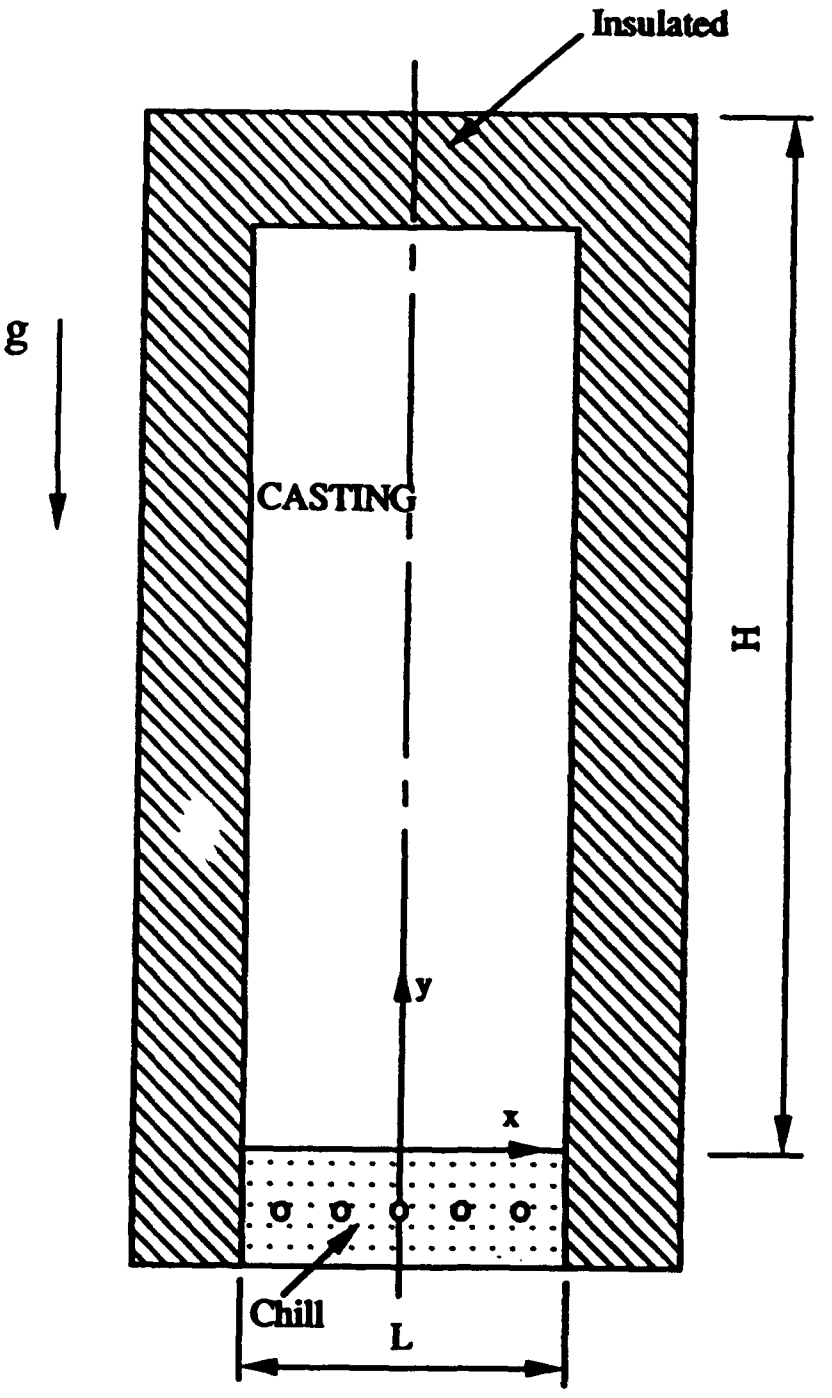


Fig. 1 Schematic Representation of a Unidirectional Solidification Chilling from Bottom.

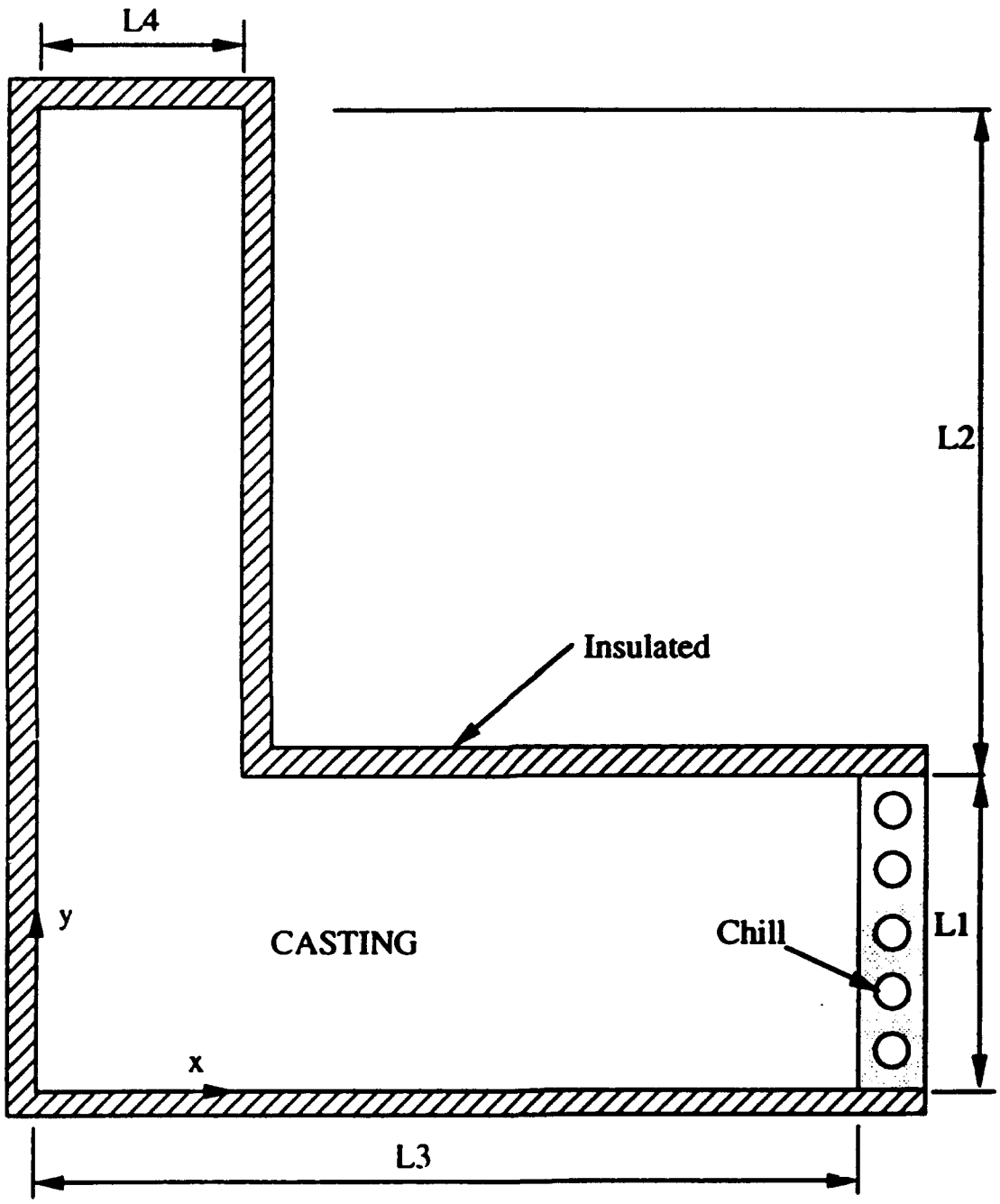


Fig. 2 Schematic Representation of a Solidification Process Chilling from Side-Wall.

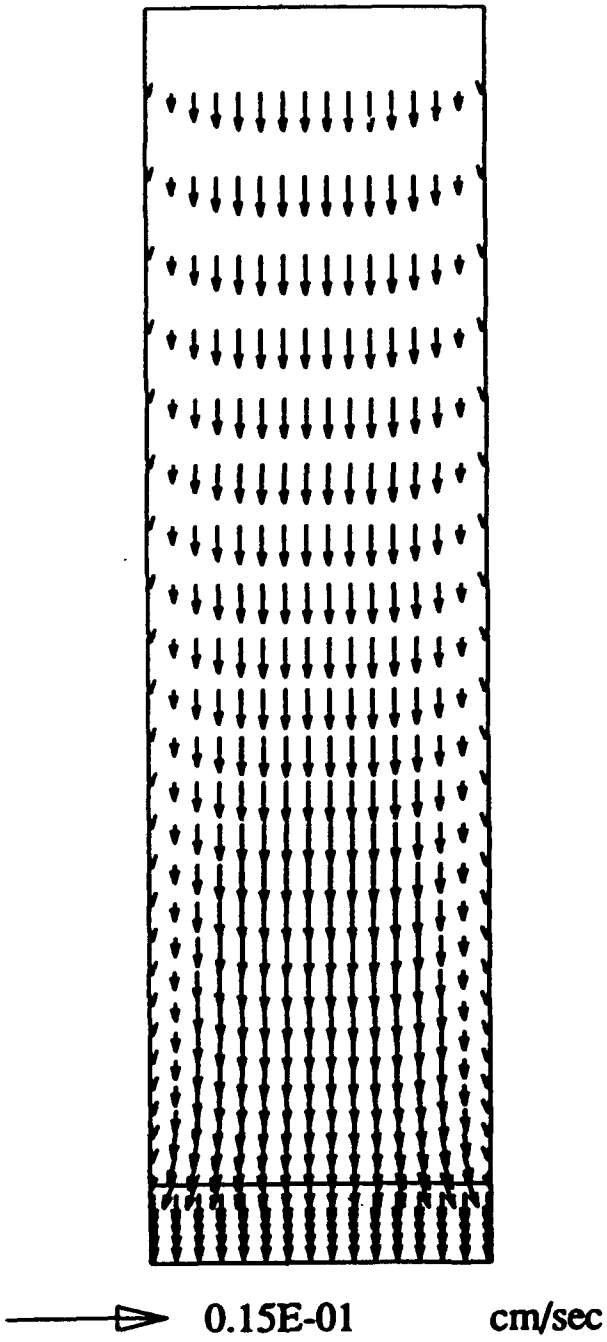


Fig. 3 Flow Pattern and Mushy Zone at Time t = 16.2 Seconds.

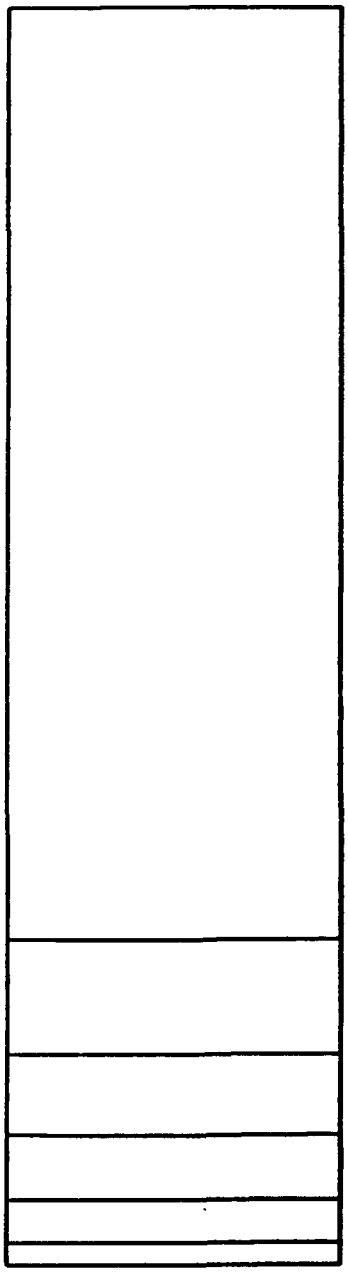


Fig. 4 Isotherms in the Casting at Time t = 16.2 Seconds.

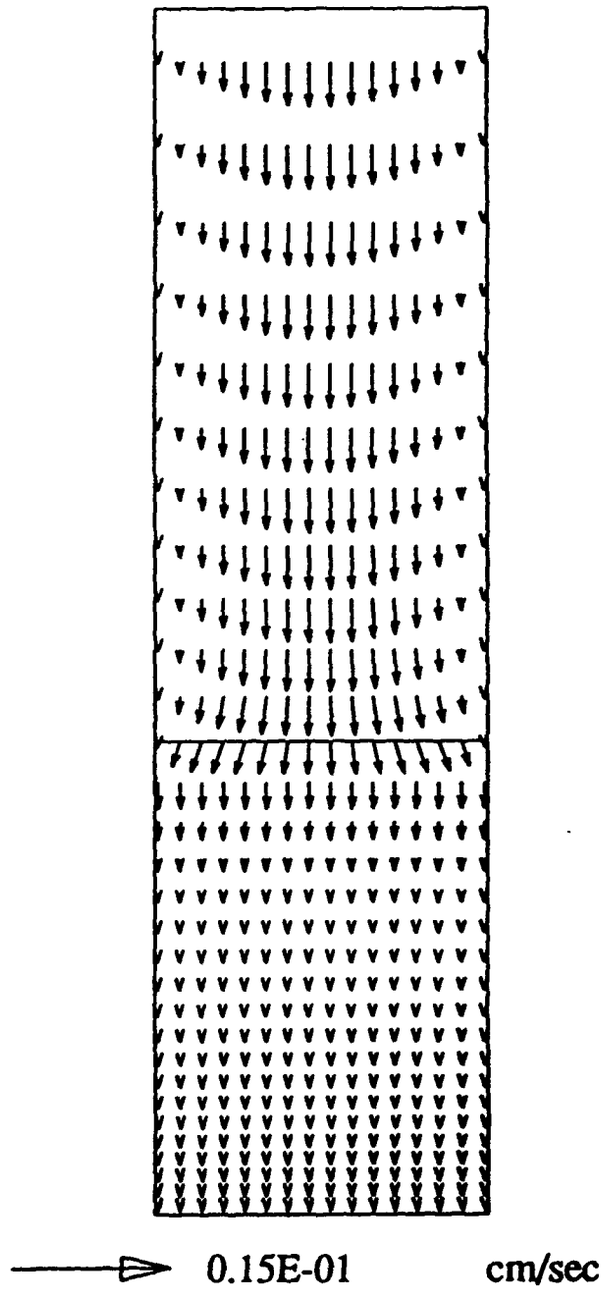


Fig. 5 Flow Pattern and Mushy Zone at Time t = 136.2 Seconds.

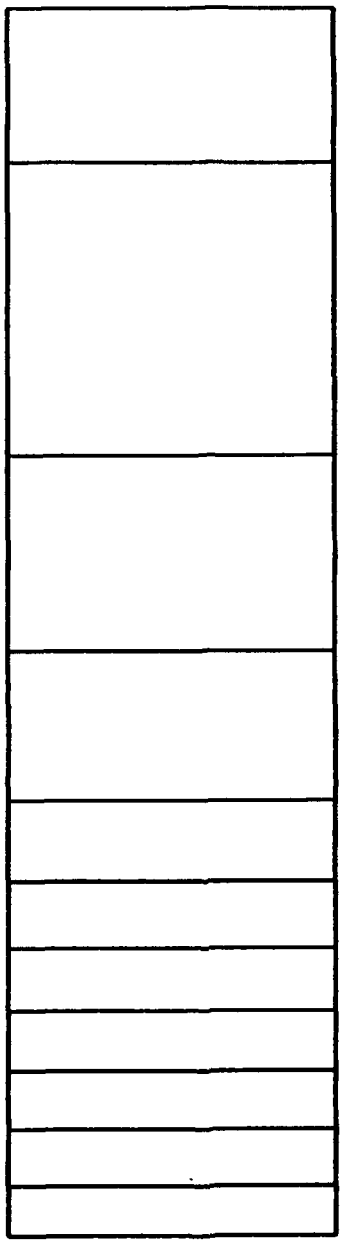


Fig. 6 Isotherms in the Casting at Time t = 136.2 Seconds.

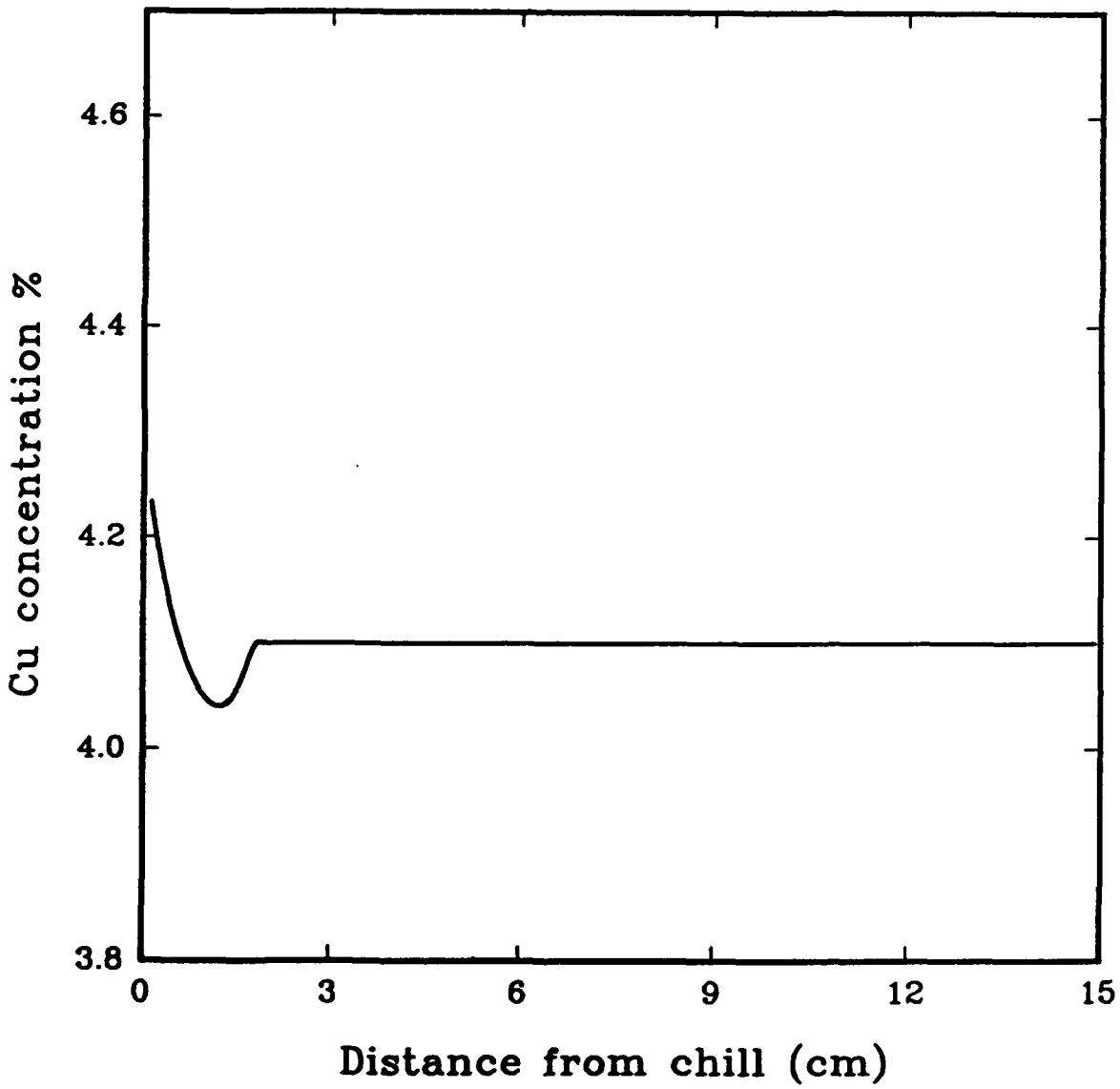


Fig. 7 Copper Concentration in the y-direction at Time t = 16.2 Seconds.

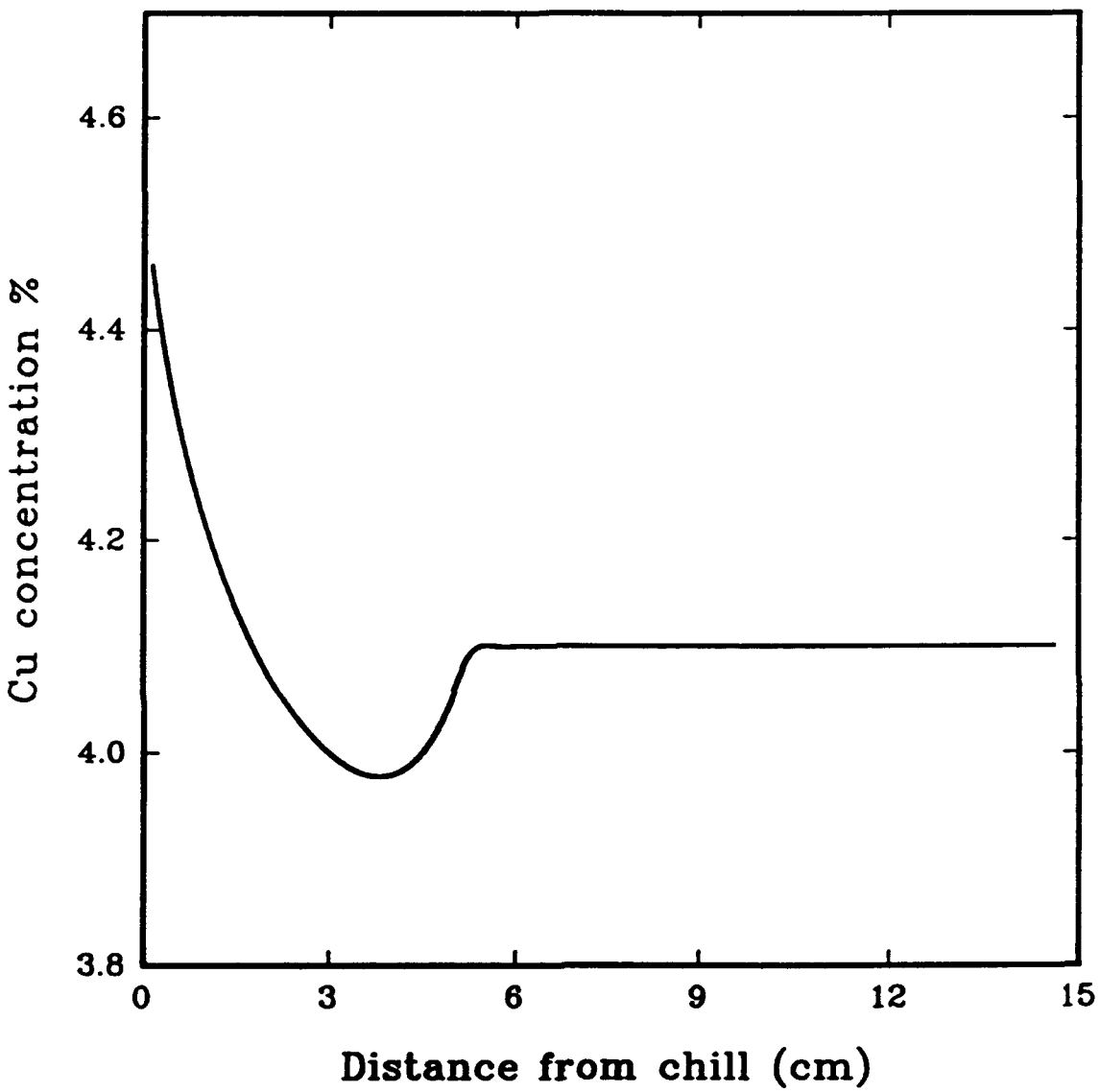


Fig. 8 Copper Concentration in the y-direction at Time t = 136.2 Seconds.

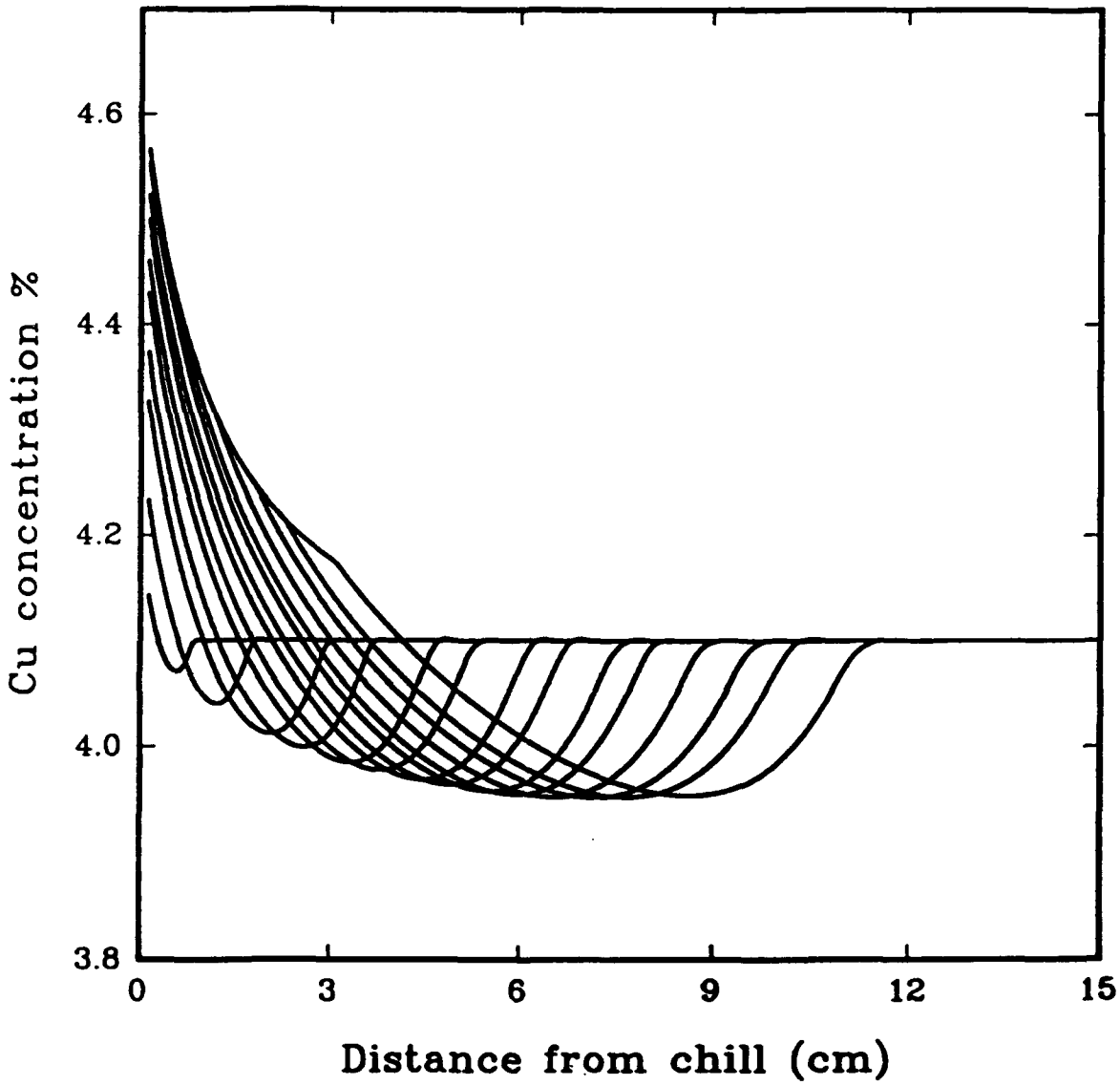


Fig. 9 Time Evolution of Copper Concentration in the Casting.

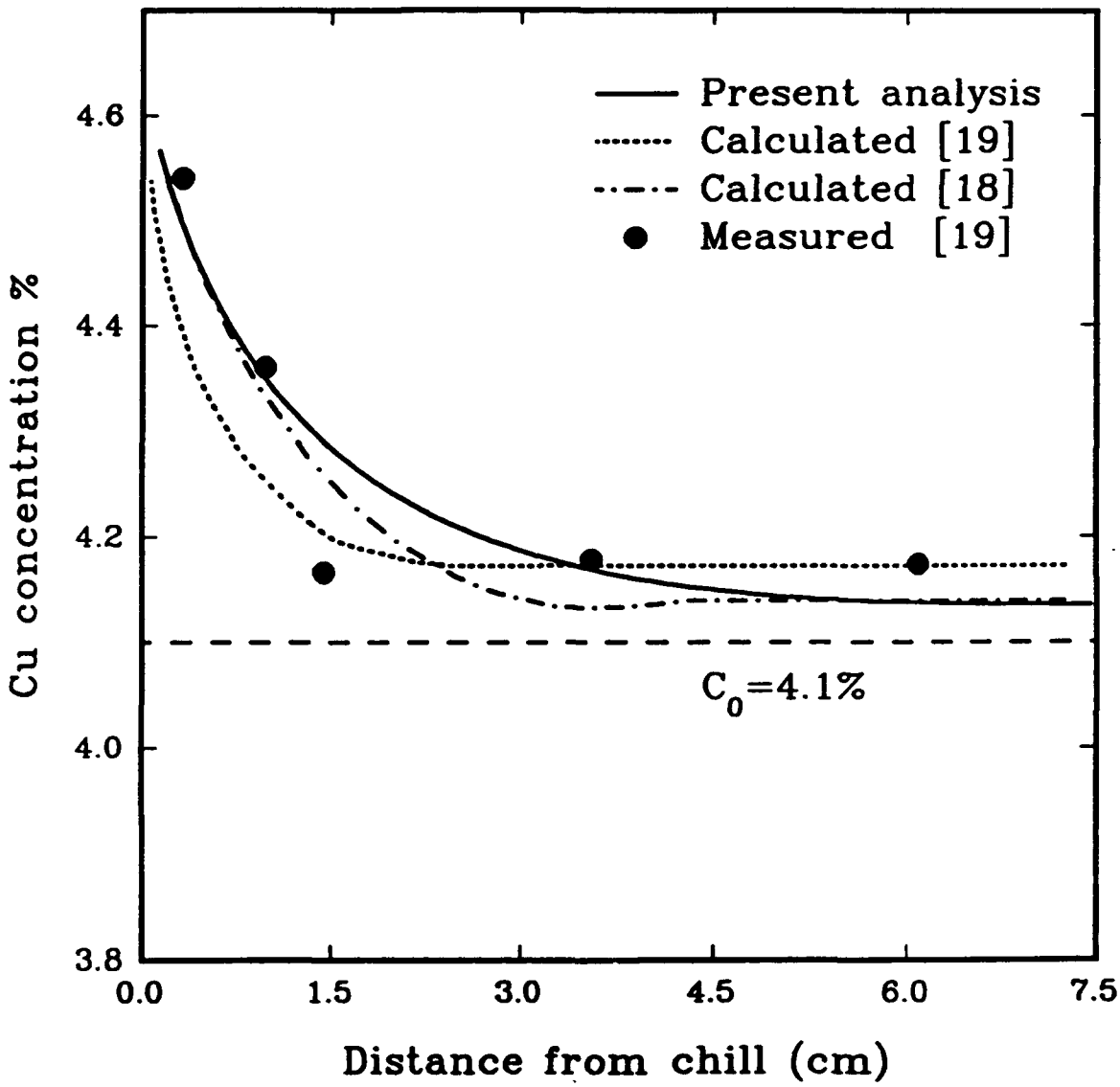


Fig. 10 Copper Concentration Distribution in the Solidified Casting.

Time = 0

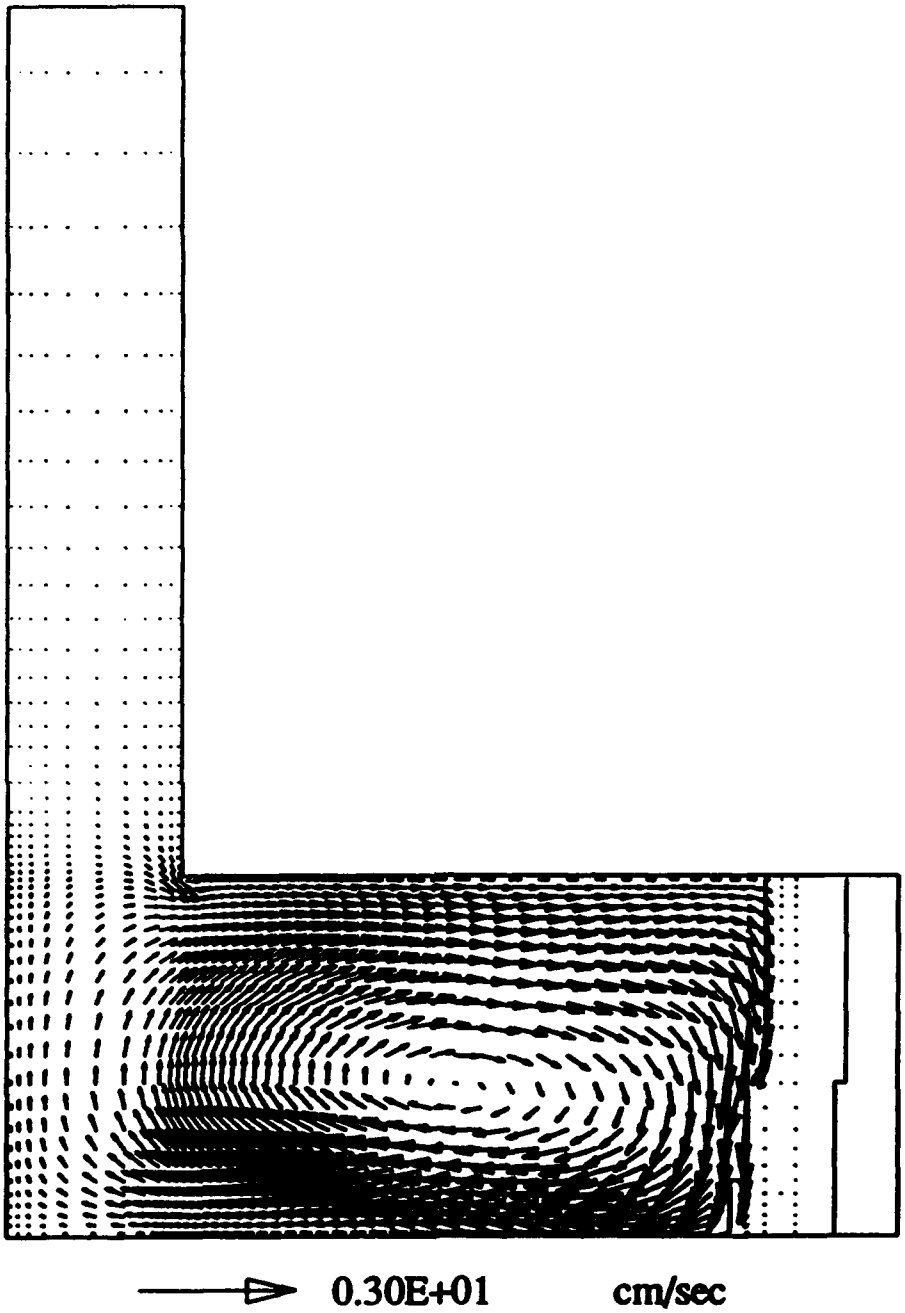


Fig. 11 Flow Pattern and Mushy Zone at Time t = 9.1 Seconds.

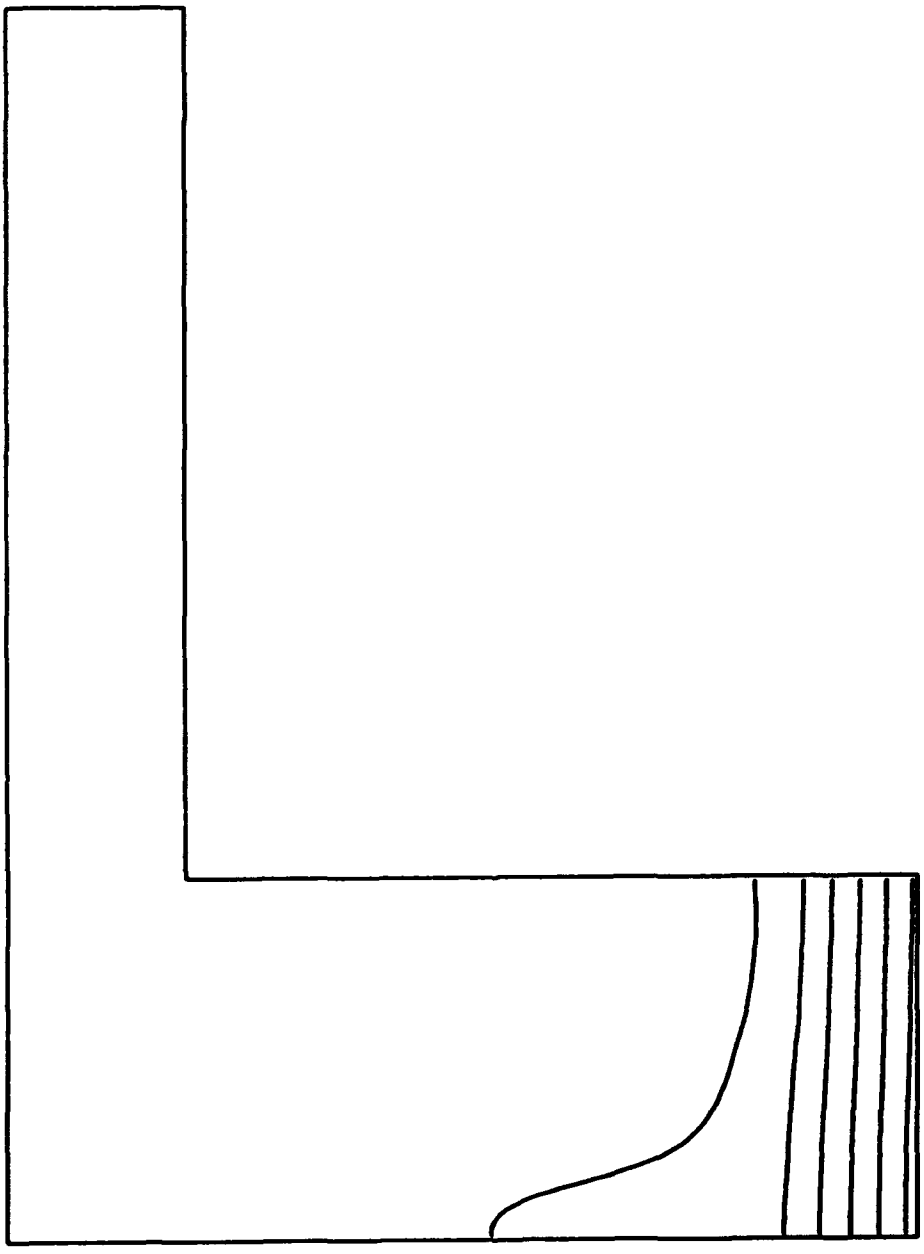


Fig. 12 Isotherms in the Casting at Time t = 9.1 Seconds.

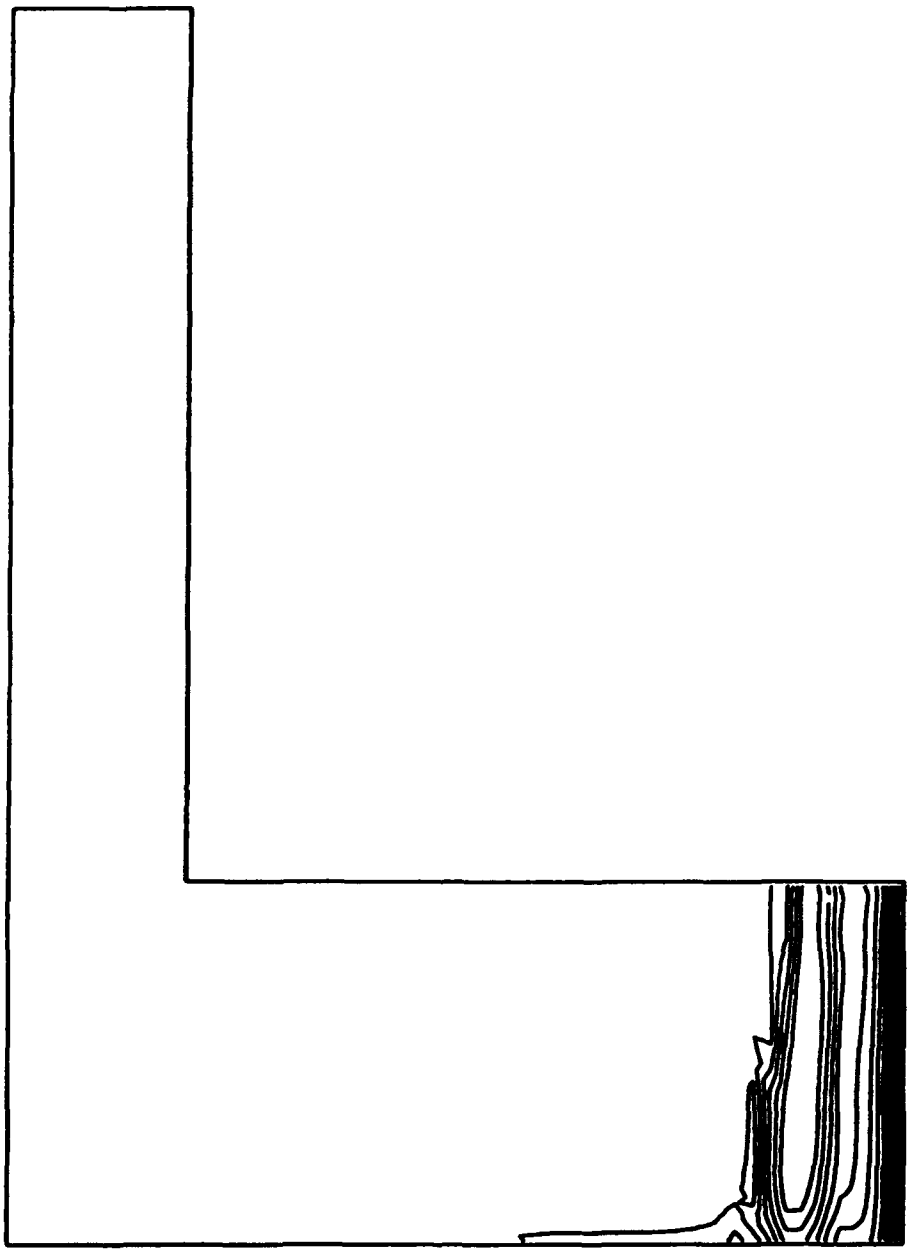


Fig. 13 Copper Concentration Profile in the Casting at Time t = 9.1 Seconds.

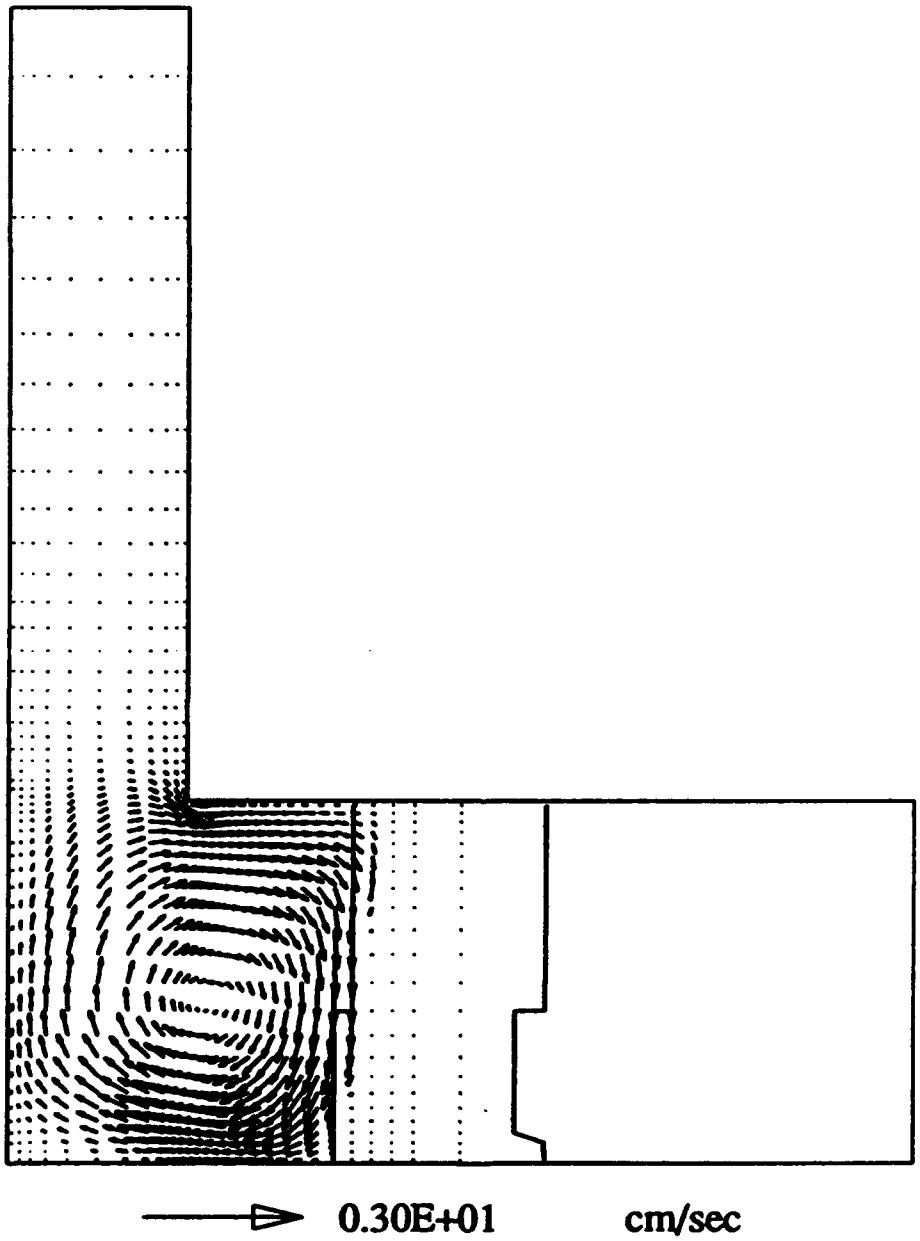


Fig. 14 Flow Pattern and Mushy Zone at Time t = 57.1 Seconds.

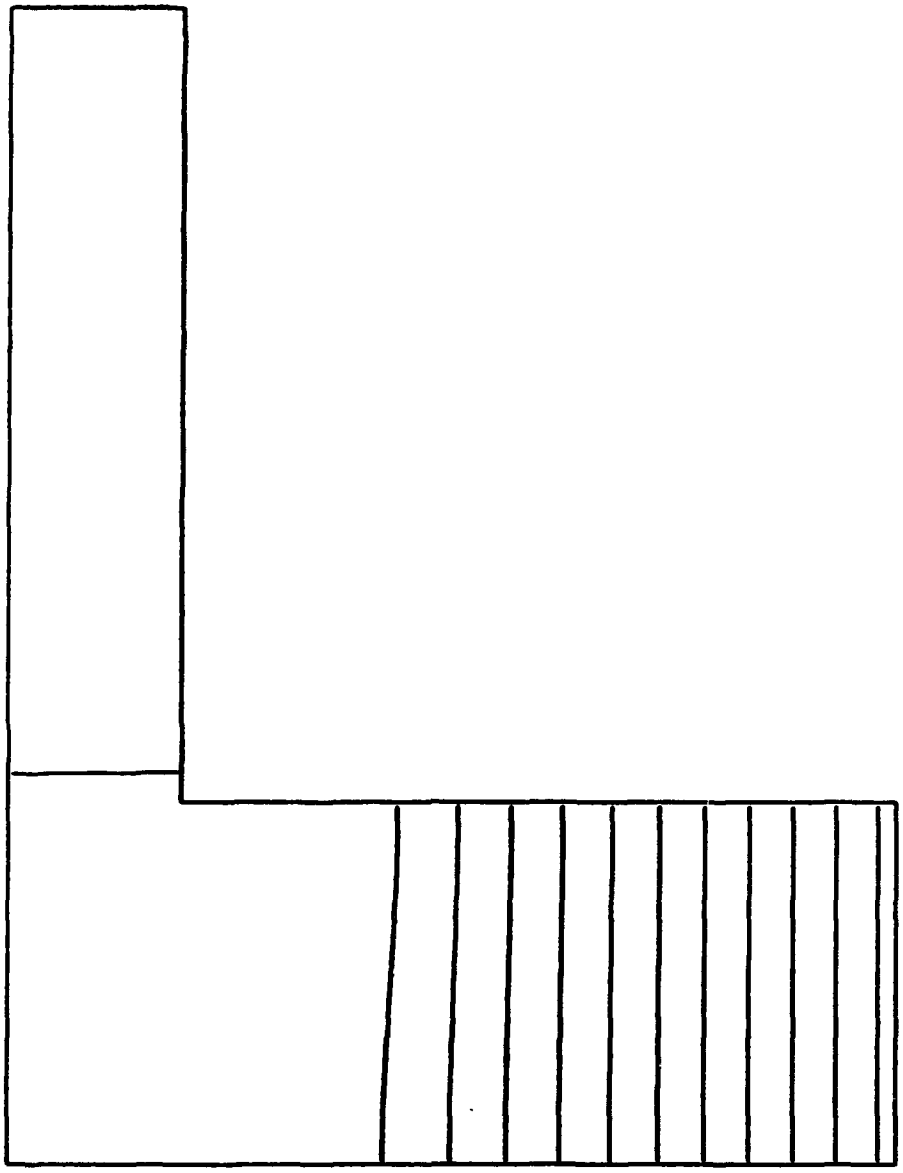


Fig. 15 Isotherms in the Casting at Time t = 57.1 Seconds.

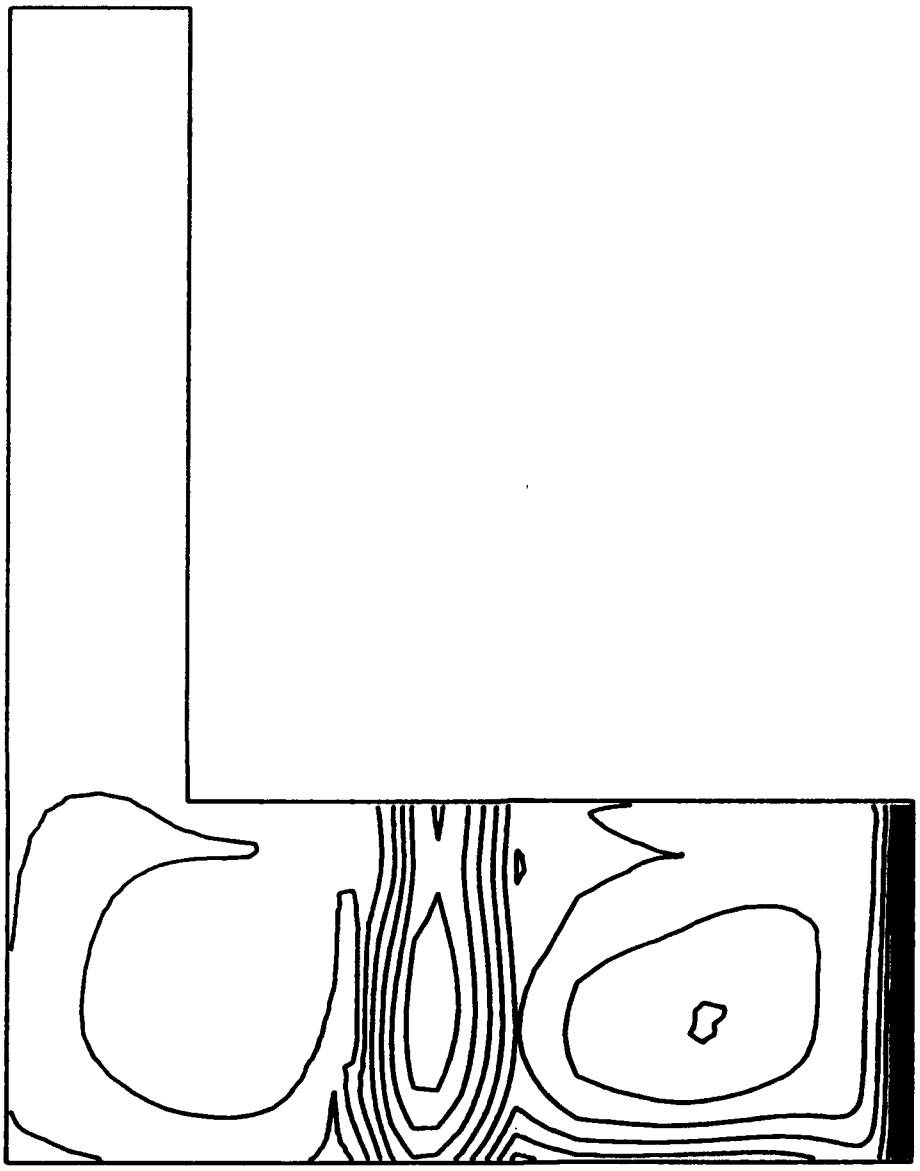


Fig. 16 Copper Concentration Profile in the Casting at Time t = 57.1 Seconds.